RANLP 2019

12th Workshop on
Building and Using Comparable Corpora

PROCEEDINGS

Edited by

Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp

5 September 2019

Proceedings of the 12th Workshop on
Building and Using Comparable Corpora, 8 September 2019 – RANLP 2019, Varna, Bulgaria

Edited by Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp

https://comparable.limsi.fr/bucc2019/

Organising Committee

- Serge Sharoff (University of Leeds, UK), Chair
- Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France)
- Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)

Programme Committee

- Ahmet Aker (University of Sheffield, UK)
- Ebrahim Ansari (Department of Computer Science and Information Technology, IASBS, Iran)
- Thierry Etchegoyhen (Vicomtech, Spain)
- Gregory Grefenstette (INRIA, Saclay, France)
- Askar Hamdulla (Xinjiang University, China)
- Hitoshi Isahara (Toyohashi University of Technology)
- Kyo Kageura (University of Tokyo, Japan)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Shervin Malmasi (Harvard Medical School, US)
- Pabitra Mitra (Indian Institute of Technology, Kharagpur, India)
- Michael Mohler (Language Computer Corp, US)
- Emmanuel Morin (Université de Nantes, France)
- Dragos Stefan Munteanu (SDL Research, Los Angeles, US)
- Lene Offersgaard (University of Copenhagen, Denmark)
- Reinhard Rapp (Magdeburg-Stendal University of Applied Sciences and University of Mainz, Germany)
- Serge Sharoff (University of Leeds, UK)
- Nasredine Semmar (CEA LIST, France)
- Michel Simard (National Research Council, Canada)
- Richard Sproat (OGI School of Science Technology, US)
- Tim Van de Cruys (IRIT-CNRS, Toulouse, France)
- Stephan Vogel (QCRI, Qatar)
- Guillaume Wisniewski (Université Paris Sud LIMSI-CNRS, Orsay, France)
- Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay, Orsay, France)

Preface – 12th BUCC at RANLP’19

In the language engineering and the linguistics communities, research on comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is primarily motivated by the need to use comparable corpora as training data for statistical Natural Language Processing applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

Comparable corpora are collections of documents that are comparable in content and form in various degrees and dimensions. This definition includes many types of parallel and non-parallel multilingual corpora, but also sets of monolingual corpora that are used for comparative purposes. Research on comparable corpora is active but used to be scattered among many workshops and conferences. The workshop series on “Building and Using Comparable Corpora” (BUCC) aims at promoting progress in this exciting emerging field by bundling its research, thereby making it more visible and giving it a better platform.

Following the eleven previous editions of the workshop which took place in Africa (LREC’08 in Marrakech), America (ACL’11 in Portland and ACL’17 in Vancouver), Asia (ACL-IJCNLP’09 in Singapore, ACL-IJCNLP’15 in Beijing, LREC’18 in Miyazaki, Japan), Europe (LREC’10 in Malta, ACL’13 in Sofia, LREC’14 in Reykjavik and LREC’16 in Portoroz) and also on the border between Asia and Europe (LREC’12 in Istanbul), this year the 12th edition of the BUCC workshop is back to Bulgaria (the first time the BUCC workshop returns to a country).

A major paradigm change in the field concerns the prevalence of Artificial Neural Networks, also appearing under the more catchy title of *Deep Learning*. Within the last five years, the Deep Learning methods shifted the balance in multilingual NLP processing towards less parallel and more comparable resources, e.g., by providing multilingual embedding spaces from monolingual corpora and by enabling Neural MT with minimal or no reliance on parallel data. Neural Networks finally make it possible to take long distance dependencies (e.g. between the words within a sentence) into account, thus overcoming a fundamental limitation of traditional n-gram-based approaches. The proceedings of this workshop present the new horizons for multilingual research with limited resources.

We would like to thank all people who in one way or another helped in making this workshop once again a success. We’re especially grateful to Ruslan Mitkov and the team of the RANLP organisers for helping us with the event.

Serge Sharoff, Pierre Zweigenbaum, Reinhard Rapp September 2019

Programme

9:20–9:30 ***Opening Remarks***

Session 1: Panel

09:30–10:30 Invited panel: Ekaterina Lapshinova-Koltunski, Sebastian Pado, Alexander Panchenko
Future directions of research in comparable corpora

Session 2: Applications of Neural Networks to Comparable Corpora

10:30–11:00 Yuri Bizzoni, Elke Teich
Analyzing variation in translation through neural semantic spaces

11:00–11:30 Maria Kunilovskaya, Shiva Taslimipoor, Tatyana Ilyushchenya
Functional Text Representations for Building Cross-Linguistic Comparable Corpora in English and Russian

11:30–12:00 ***Coffee Break***

Session 3: Inducing Lexicons from Comparable Corpora

12:00–12:30 Ebrahim Ansari, M.H. Sadreddini, Mahsa Radinmehr and Ziba Khosravan
Extracting Bilingual Persian Italian Lexicon from Comparable Corpora Using Different Seed Dictionaries

12:30–13:00 Kim Steyaert and Ayla Rigouts Terryn
Multilingual Term Extraction from Comparable Corpora: Informativeness of Monolingual Term Extraction Features

13:00–14:30 ***Lunch Break***

Session 4: Building Comparable corpora

14:30–15:00 Benjamin K. Tsou and Kapo Chow
From the cultivation of comparable corpora to harvesting from them: A quantitative and qualitative exploration

Session 5: Shared Tasks

15:00–16:00 Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff
Towards the Fourth BUCC Shared Task: inducing bilingual lexicons through comparable corpora

16:00–16:15 *Closing remarks*

Table of Contents

Analyzing variation in translation through neural semantic spaces

Yuri Bizzoni
University of Saarland, Germany
yuri.bizzoni@uni-saarland.de

Elke Teich
University of Saarland, Germany
e.teich@mx.uni-saarland.de

Abstract

We present an approach for exploring the lexical choice patterns in translation on the basis of word embeddings. Specifically, we are interested in variation in translation according to translation mode, i.e. (written) translation vs. (simultaneous) interpreting. While it might seem obvious that the outputs of the two translation modes differ, there are hardly any accounts of the summative linguistic effects of one vs. the other. To explore such effects at the lexical level, we propose a data-driven approach: using neural word embeddings (Word2Vec), we compare the bilingual semantic spaces emanating from source-to-translation and source-to-interpreting.

1 Introduction and Related Work

Our research question stems from the field of translation studies. Revisiting the notion of 'translationese' (Gellerstam, 1986), i.e. the specific linguistic traces left in the translation product by the process of translation, we are interested in patterns of lexical choice in translation versus interpreting. To explore this, we need (a) summaries of the dominant lexical choices made in translation and interpreting and (b) a method of comparing them.

Existing research on translationese (Baker, 1996) has mainly focused on (sets of) predefined features (e.g. type-token ratio, sentence length, part-of-speech distributions), applied in classification tasks comparing translations and original texts (Baroni and Bernardini, 2006; Volansky et al., 2015; Rubino et al., 2016). While this work has brought genuine insights regarding the language of translation, we still have a fairly fragmented picture of translation behavior and its many facets (Lapshinova-Koltunski, 2013, 2015). For instance, it has been shown that translations exhibit source language interference or shining-through (Toury, 1995; Teich, 2003), against the assumption of translation universals; or that certain groups of translators show higher convergence in translation choice than others (see e.g. (Martínez Martínez and Teich, 2017) who study the outputs of translation learners and professionals by entropy).

Here, we are interested in written translation vs. simultaneous interpreting. Among the known differences are more frequent and different kinds of omissions in interpreting (He et al., 2016) and, depending on the source - target language pair, unusual word orders (Collard et al., 2018). However, there is no comprehensive, systematic picture yet, also due to the fact that specific and systematic studies of interpretation are a relatively recent phenomenon (Pöchhacker, 2016).

Nonetheless, we can formulate some hypotheses. Beyond the notorious difficulties of bridging a "message" between two languages - difficulties that are constantly analyzed in translation studies (Eades, 2011; Li, 2019) - the process of interpreting is complicated by the dire time constrains of the process and by the absence of an editing phase, essential in many translation processes (Schaeffer et al., 2019). We might thus assume that due to high cognitive pressure, interpreters may not be able to adapt their output to the target language norms as well as translators do, which might be reflected in lower lexical richness and lexically less coherent interpreting output compared to translation output.

On the computational side, the approach proposed here is related to attempts at making word embeddings fruitful for linguistic analysis, notably modeling diachronic language change (Dubossarsky et al., 2017; Fankhauser and Kupietz, 2017; Bizzoni et al., 2019). Also, there is some resemblance to the problem of creating domain-specific word embeddings (Zhang et al., 2019; Wang et al., 2018).

Concretely, the method we propose here aims at building bilingual word embeddings from aligned corpora. In the last years, a significant amount of research has gone into the construction of more effective multilingual word embeddings (Zhang et al., 2017; Artetxe et al., 2018) from smaller datasets (Artetxe et al., 2017) or with the help of multimodal data (Singhal et al., 2019).

But while most works on multilingual distributional semantics focus on creating consistent spaces (Huang et al., 2018) showing robust properties across languages (Brychcín et al., 2019), our aim is creating semantic spaces that model the lexical choices of a *specific* kind of linguistic behavior, i.e. translation, which we call here *translation spaces*. Specifically, we train two bilingual distributional models on two monolingually comparable corpora, a larger one of translation, and a smaller one of interpreting, and we compare them to detect differential patterns of translation mode-induced lexical choice.

It is important to underline that in this first stage, the gist of our analysis comparing semantic spaces will be qualitative (Sections 4.1-4.3). Qualitative analyses are somewhat easier on bilingual than on monolingual spaces, for the reason that in bilingual spaces we often know the "ground truth" (e.g. we know that the Spanish translation of *Germany* is *Alemania*), while the similarities displayed by monolingual word embeddings are harder to judge case by case. Therefore, our bilingual word embeddings are directly comparable and we are able to present a conclusive quantitative perspective on the spaces' overall topology as well (Section 4.4). In that case, we will just consider the mean distances, without looking at the actual words in a cluster.

For all our experiments we used gensim's implementation of Word2Vec (Mikolov et al., 2013).

2 Corpora

Our data set is composed of Spanish and German translations of the same English source (speeches from the European Parliament) (Karakanta et al., 2018) as well as interpreted speeches in the same two target languages. For written translation, each language is represented by circa 20 million characters and 130.000 sentences. For German, we have created a corpus of interpreted speech of English into German from the European Parliament including materials from existing corpora (Sandrelli and Bendazzoli, 2005; Bernardini and Milievi, 2016). The resulting interpreting corpus is strictly comparable to the translation corpus in terms of register and domain but contains much less material (568.230 characters and 3.397 sentences per language).

3 Methodology

3.1 Creating translation spaces

The input for a translation space is constituted by the tokenized, concatenated aligned sentences of a source-translation corpus. In other words, each sentence from a source text X is concatenated with its translation in a target text Y, creating a bilingual pseudo-sentence. If we were dealing with an idealized word-by-word translation, this pseudo-sentence would be simply composed by lexical source-target pairs; in the case of a more realistic translation, we can still confidently expect that a percentage of the words in the source language will find a direct target correspondence within the same pseudo-sentence. After creating the pseudo-sentences, we train a standard skip-gram Word2Vec model on them, using as context window the mean + standard deviation length of the sentences (in our case, we set each word's context at 160 words, which is the double of the mean sentence length plus standard deviation). Before training, the words in each aligned sentence were shuffled: this proved to yield slightly better results.

The logic of this approach is that words having a consistent translation in an aligned corpus will share very similar contexts, ending up in close proximity in the resulting distributional space.

An important problem to address is the variability of Word2Vec's results at different run times. While the specific cosine similarity is bound to undergo oscillations between different runs, all the rankings we present in the following tables have been verified through multiple runs: in other words, if the cosine similarities slightly changed, the ordering and the magnitude of the results remained the same. In future we intend to verify the stability of our spaces more consistently (see Section 5).

3.2 Probing the translation spaces

Words that translate each other in a very consistent way throughout the corpus appear to be very close in the resulting semantic space, and are often each other's nearest neighbours. For example, in the English-Spanish translation space, the nearest neighbour of *Germany* is *Alemania*, of *Italy Italia*, and so forth. Also, country names in both languages create a tight semantic cluster, and happen to be in the same lexical neighborhood (see Table 1).

This example shows the qualities of a space deriving from a translation where each term has one and one only direct equivalent in the other language: the group of country names forms a cluster which is both bilingually sound (*Alemania* is the closest word to *Germany*) and semantically coherent (*Italy*, *Italia* and *France* are the three following neighbours of *Alemania*). We can consider this cluster as representative of extremely faithful translations: situations where each word in X has its undiscussed equivalent in Y. Such peculiar cases of "extreme" source text fidelity guarantee both semantically and translationally sound distances. On the opposite side of the spectrum, we can find elements that rarely have a single, obvious equivalent in another language: function words. Words like Spanish *el* or English *to* and *if*, do not have a meaningful closest neighbour in the other language and are on average further apart from other words than words with one obvious translation: they form looser clusters.

Translation spaces are of particular interest in the cases between these two extremes. Both the identity and the distance of neighbours become indicative of a translation "style". For example, in the same space, we find that *war* is closest to *guerra* (cosine similarity .91) but also relatively close to *terror* (0.71), *fria* (0.69), *cold* (0.69): *war* seems to be consistently translated, and in a semantically quite coherent cluster. The nearest neighbour of *voz* is *voice*, with a cosine similarity of 0.91, but its second nearest neighbour, *solidaridad*, has a similarity of only 0.57, followed by words mainly in Spanish, such as *sola* and *expresarse*; *voz* has a consistent translation, but belongs to a less obvious paradigm.

The comparison with the country names, where each word is nearest to its translation and very near to other country names in both languages, is helpful to see how we are moving towards more semantically complex cases: *war* and *voz* are words that have a preferential translation, but do not belong to conventionalized paradigms as predictably translated as country names.

Germany	**war**	**voz**
Alemania.95	guerra.91	voice.91
Italia.86	terror.71	solidaridad.57
Italy.86	fria.69	sola.56

Table 1: Three words and their three nearest neighbours with cosine similarities in the Spanish-English translation space.

If instead a word is *not* consistently translated, there are two possible configurations in the translation space:

1. The word is close to its various translations in the space, but the similarity is relatively low. This seems to represent the case of well defined polysemy, where one word is consistently translated with one among N choices in the target language: for example *fear* is close to *temo*, *miedo* and *temor*, and their cosine similarities are between 0.62 and 0.7.

2. The word isn't close to any translating term in the other language, and does not present a high similarity to its neighbours. This seems to represent the case of words that are particularly hard or impossible to translate with one term in the target language. Such words produce many contextual translations, rephrasing, or omissions, and this "'productivity" in turn makes their distributional profile relatively idiosyncratic, distancing them from all other points in the space. For example, *somehow* has no close neighbours in Spanish, and its nearest term in the space, *foolish*, has a cosine similarity of only 0.49; the nearest neighbour of *weekend* is *week* (0.57) and the nearest neighbour of *insight* is *spirit* (0.41).

This simple mirroring between the source fidelity of a translation and the tightness of a distributional cluster can be a special way to detect several translation behaviours (see Table 2).

4 Translation spaces for comparable corpora

As a use case, we want to adopt this system of building distributional spaces to compare lexical

gentes	**the**	**palestinian**
oppressed.54	mandato.23	palestino.9
gente.52	de.23	palestinos.88
pueblo.52	cachemir.22	israeli.87
population	**quiero**	**sucesor**
inhabitants.61	quisiera.88	successor.84
viven.57	deseo.75	gallant.6
living.57	desearia.7	franco.56

Table 2: Words with no direct translation in loose clusters (*gentes*, *the*), words with direct translation in tight semantic clusters (*palestinian*), words with some semantic tightness but no direct translation (*population*, *quiero*), words with direct translation in a loose semantic cluster (*sucesor*). In the majority of cases, no direct translation means lower semantic similarity with the nearest word, ergo looser clusters. An "untranslatable", be it real or perceived, doesn't have single words that share its context with the same regularity of a translating term, and thus tendentially creates looser groups.

fidelity between the translation corpus and a comparable interpreting corpus.

We conduct first a qualitative analysis of the differences, and then a quantitative analysis of the topological differences between the two spaces.

4.1 English-German translation space

Following the procedure described in the previous section, we train a translation space on the tokenized and aligned sentences of the English-to-German written translation corpus, with a context window of 160 words and a dimensionality of 300.

This space seems to behave coherently with what we would expect:

1. Words with single, highly preferred translations form translation and semantic tight groups: *Germany* is close to *Deutschland* (0.94) and *Belgien* (0.84); *Mord* is close to *murder* (0.95) and *brutale* (0.86). Technical terms too tend to display high nearest neighbour similarities: *unemployment* - *Arbeitslosigkeit* (0.89), *decriminalisation* - *Entkriminalisierung* (0.96).

2. Words belonging to semantically complex paradigms fall relatively close to their preferential translation when they have one, but their clusters are looser: *force* is the nearest neighbour of *Kraft* (0.67) and *Friedenstruppe* (0.64).

3. Words with various translations fall close to their equivalents, but their similarities are low: *happy* is close to *glücklich* (0.62), *erfreut* (0.52), *zufrieden* (0.51).

4. Words without a single term translation are at the center of very loose clusters, with nearest neighbours' similarities ranging between 0.6 and 0.4.

Both in this space and the English-Spanish one, geometric analogies of the sort of "'man : woman = king : x" (Mikolov et al., 2013) are possible with various terms: "man : woman = Mann : x" returns *Frau*; "glücklich : sad = happy : " returns *traurige*; "Freiheit : Presse = freedom : x" returns *press* and *newspapers*. In other words, the sum vector of *Freiheit* + *freedom* minus *Presse* returns a point that is closest to *press*.

While such results are the effects of consistent translation embeddings, this particular space also shows peculiarities that are due to the specifics of German compounding: the sum vector of *freedom* + *press* is close to *pressefreiheit* (0.70); the sum vector of *freedom* + *expression* is closest to *meinungsfreiheit* (0.88), and so forth.

Interestingly, the closest neighbours of *meinungsfreiheit* are *expression* and *freedom* with relatively high degrees of similarity (0.88 and 0.79): terms that have a multi-word consistent translation can still exhibit tight clustering properties.

4.2 English-German interpreting space

The interpreting space shows properties in common with the translation space, but with relevant differences.

1. Words with a highly preferred single translation fall closest to such translation, but do not seem to form semantically cohesive clusters: *Germany* is closest to *Deutschland* (0.98), but is not in a cluster of country names.

2. Words that showed a variety of translation neighbours in the translation space either present a single, very close meaningful neighbour (*zufrieden* has a 0.86 similarity to *satisfied*, but no other English words appear in its immediate vicinity), or tend to show no meaningful clustering at all.

Full Translation Space	Small Translation Space	Interpreting Space
Germany: Deutschland.94, Belgien.84, Frankreich.84	**Germany:** Deutschland.99, vivendi.84, France.84	**Germany:** Deutschland.98, politically.8, tragbar.78
somehow: irgendwie.7, Wahrheit.6, erwecken .59,	**somehow:** irgendwie.84, anhängen.83, pollute.8	**somehow:** enjoy.64, speaks.63, volumes.63
happy: glücklich.63, erfreut.53, zufrieden.51	**happy:** glücklich.88, verspatung.74, soweit.72	**happy:** glad.67, glücklich.65, m.65
Vertrag: treaty.79, Nizza.77, Nice.72	**Vertrag:** treaty.89, idea.66, settle.65	**Vertrag:** treaty.93, Lissabon.84, Lisbon.8

Table 3: Three nearest neighbours of *Germany*, *somehow*, *happy* and *Vertrag* for the full scale Translation Space, the down-sampled Translation Space, and the Interpreting Space.

3. Finally, words with no direct translation remain inside loose, sparse clusters.

4.3 Subsampling and comparison

The main problem with comparing the two spaces is difference in corpus size, the translation corpus being significantly larger than the interpreting corpus.

To take this aspect into account, we randomly sampled the translation corpus in order for it to have the same number of aligned sentences as the interpreting corpus, and trained a new distributional space on it.

A qualitative presentation of the difference between the three spaces is in Table 3.

Four main observation can be made:

1. The down-sampled translation model keeps some looser semantic cohesion with country names, while the interpreting model seems able to "only" retrieve the direct term translation;

2. Adverbs such as *almost*, *probably*, *irgendwie* etc. retrieve an equivalent in the translation spaces, but not in the interpreting space.

3. In some cases, such as in the case of *happy*, the effect of data scarcity is that of strengthening the relation between a term and one of its possible translations, probably due to the absence of alternatives in the down-sampled corpus; this makes the loose similarity of *happy* with *glücklich* in the interpreting corpus more relevant.

4. The relation between translatability and cosine similarity seems to hold through the spaces: if two neighbours translate each other, their similarity tends to be higher than if they are simply semantically related.

4.4 Topological comparison of the spaces

Given these observations, we can proceed to a comparison of some topological properties of the translation sub-sampled space and the interpreting space (see Table 4 for a summary).

We note that despite being of equal length, the translation model has more words than the interpreting model: 18 592 versus 10 524. For this comparison, we will focus on the 6 753 words that they have in common.

The average word similarity within the translation model is 0.26, six points higher than the average similarity within the interpreting model. But if we limit our computation to every word's nearest neighbour in each model, we see a different picture emerging. The distance between the models shrinks to no true significance, with the interpreting space showing even a slightly higher similarity than the translation: nearest neighbours in the translation space have an average cosine similarity of 0.85, those in the interpreting space of 0.86.

The average word similarity is different between the two spaces, but the nearest neighbour similarity is approximately the same. In other words, interpreting spaces present a less homogeneous distribution than translation spaces: they display words with nearer neighbours in looser clusters. This distribution seems to go along with our observations of a two-folded fidelity to the source: interpreting seems to show a high level of fidelity with respect to unambiguous, domain-specific words (*treaty* - *Vertrag*, *president* - *Präsident*) where the translation space presents a lower degree of similarity (more diversity in the translation). The result of these cases is close nearest neighbours followed by lower similarity words in interpreting spaces; and more distant nearest neighbour followed by closer alternatives in translation spaces.

At the same time, other categories of words, such as adverbs (*irgendwie*, *really*, *next*), and some non-domain specific words (*tag*, *muss*) seem to have no systematic equivalent in the interpreting space, while they do retrieve a translating nearest neighbour in the translation spaces, both full and "down-sampled". This may suggest a preference on the part of the interpreter for a precise translation of domain-specific content at the expense of interpersonal or textual expressions. These opposed tendencies could be the cause of the special topology we seem to observe in the spaces.

	Translation	**Interpreting**
vocab size	18592	10524
avg simil.	0.268	0.213
1st neigh.	0.851	0.860
10th neigh.	0.723	0.685

Table 4: Vocabulary size, average overall similarity, first and tenth nearest neighbour average similarity for the down-sampled translation space and the interpreting space. The mean difference between the first and tenth neighbour also shows the "loosening" of the similarity queues in the interpreting space, symptom of generally looser word clusters.

5 Conclusions and Future Work

We have presented a method of exploring variation in translation, here focusing on translation vs. interpreting, using neural word embeddings. Creating two models, a source-translation model and a source-interpreting model, from the same domain (European Parliament speeches) we explored similarities and differences between the two lexical-semantic spaces. To obtain better comparability, we down-sampled the dimensions of the translation corpus in order to avoid mistaking frequency effects for true translation behaviours.

Our comparison has revealed both differences in the overall topology of two semantic spaces (looser word clusters in interpreting compared to translation) as well as differences in how translators vs. interpreters handle certain types of vocabulary (e.g. domain-specific vs. general words). We can speculate on some possible reasons of the differences between the spaces: for example, the two-folded fidelity of interpretation could be due to the fact that while interpreting forces a more deliberate rephrasing of the source (which also comes with the apparent sacrifice of interpersonal or textual expressions), formulaic or highly predictable words are easier to translate always with the same equivalent. Nonetheless, we find that more research has to be done in order to make such claims substantial.

In our ongoing work, we use the same method for looking at other variables, e.g. the influence of source language on the translation output and the level of translation *expertise* (learner vs. professional), and analyze translation spaces further in terms of entropy, as an index of lexical variation. Another matter we want to address more consistently is that of Word2Vec's possible sensitivity to words' frequency. We think that our spaces are more resistant than monolingual spaces to random initialization simply because they are modelling a more clear-cut phenomenon: if a low frequency word has a consistent translation, its distributional profile will still be uniquely similar to that of the translation. Nonetheless, we intend to evaluate this method more substantially, comparing the spaces' results to bilingual dictionaries and synthetic data, which could also help us assess the impact of frequency effects. Also, we intend to compare this method's results with the results of a post-training aligned bilingual space, and to use the proposed method for translation evaluation, complementing it with other means of comparative textual analysis, such as relative entropy.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297* .

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers, editor, *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, John Benjamins, Amsterdam and Philadelphia.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3):259–274.

S. A. Ferraresi Bernardini and M. Milievi. 2016. From epic to eptic exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28:61–86.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Katrin Menzel, Pauline Krielke, and Elke Teich. 2019. Grammar and meaning: Analysing the topology of diachronic word embeddings. In *Proceedings of First Workshop on Computational Approaches to Historical Language Change, ACL*.

Tomáš Brychcín, Stephen Taylor, and Lukáš Svoboda. 2019. Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Systems with Applications* .

Camille Collard, Heike Przybyl, and Bart Defrancq. 2018. Interpreting into an s.o.v. language: memory and the position of the verb, a corpus-based comparative study of interpreted and non-mediated speech. *Meta* 63(3):695–716.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. pages 1136–1145.

Domenyk Eades. 2011. Translating english modal expressions: an arab translator trainees perspective. *Babel* 57(3):283–304.

Peter Fankhauser and Marc Kupietz. 2017. Visualizing language change in a corpus of contemporary german .

Martin Gellerstam. 1986. *Translationese in Swedish novels translated from English*. CWK Gleerup.

He He, Jordan Boyd-Graber, Jordan Boyd Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of NAACL-HLT*. San Diego, CA, pages 971–976.

Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Multi-lingual common semantic space construction via cluster-consistent word embedding. *arXiv preprint arXiv:1804.07875* .

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. Preserving metadata from parliamentary debates. In Darja Fier, Maria Eskevich, and Franciska de Jong, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Ekaterina Lapshinova-Koltunski. 2013. VARTRA: a comparable corpus for analysis of translation variation. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*. Sofia, Bulgaria, pages 77–86.

Ekaterina Lapshinova-Koltunski. 2015. Variation in translation: Evidence from corpora. *New directions in corpus-based translation studies* pages 93–113.

Saihong Li. 2019. A corpus-based multimodal approach to the translation of restaurant menus. *Perspectives* 27(1):1–19.

José Manuel Martínez Martínez and Elke Teich. 2017. Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In Larisa Cercel, Marco Agnetta, and María Tereza Amido Lozano, editors, *Kreativitt und Hermeneutik in der Translation*, Narr, Tbingen, pages 103–126.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Franz Pöchhacker. 2016. *Introducing interpreting studies*. Routledge.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of NAACL*. San Diego, CA, pages 960–970.

Annalisa Sandrelli and Claudio Bendazzoli. 2005. Lexical patterns in simultaneous interpreting: A preliminary investigation of epic (european parliament interpreting corpus). In *Proceedings of the Corpus Linguistics Conference Series 1*.

Moritz Schaeffer, Anke Tardel, Sascha Hofmann, and Silvia Hansen-Schirra. 2019. Cognitive effort and efficiency in translation revision. In *Quality Assurance and Assessment Practices in Translation and Interpreting*, IGI Global, pages 226–243.

Karan Singhal, Karthik Raman, and Balder ten Cate. 2019. Learning multilingual word embeddings using image-text data. *arXiv preprint arXiv:1905.12260* .

Elke Teich. 2003. *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Gideon Toury. 1995. *Descriptive Translation Studies and beyond*. John Benjamins, Amsterdam/Philadelphia. https://doi.org/10.1075/btl.4.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30(1):98–118. https://doi.org/10.1093/llc/fqt031.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics* 87:12–20.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 1959–1970.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data* 6(1):52.

Multilingual Term Extraction from Comparable Corpora: Informativeness of Monolingual Term Extraction Features

Kim Steyaert and Ayla Rigouts Terryn

LT^3 Language and Translation Technology Team
Ghent University, Groot-Brittanniëlaan 45, 9000 Gent;
ayla.rigoutsterryn@ugent.be; kim.steyaert@gmail.com

Abstract

Most research on bilingual automatic term extraction (ATE) from comparable corpora focuses on both components of the task separately, i.e. monolingual automatic term extraction and finding equivalent pairs cross-lingually. The latter usually relies on context vectors and is notoriously inaccurate for infrequent terms. The aim of this pilot study is to investigate whether using information gathered for the former might be beneficial for the cross-lingual linking as well, thereby illustrating the potential of a more holistic approach to ATE from comparable corpora with re-use of information across the components. To test this hypothesis, an existing dataset was expanded, which covers three languages and four domains. A supervised binary classifier is shown to achieve robust performance, with stable results across languages and domains.

1 Introduction

Bilingual automatic term extraction (ATE) from comparable corpora aims to identify equivalent term pairs cross-lingually in monolingual corpora that are similar in terms of size, topic and style. Strongly related to bilingual lexicon induction (BLI), the main difference is that ATE from comparable corpora focuses on terminology (domain-specific, specialised vocabulary), rather than general language. Despite the difficulty of finding cross-lingual equivalents in unaligned text, comparable corpora have a substantial added value over parallel corpora, since they are much easier to create and, therefore, less expensive. This is especially relevant for low-resourced languages and specialised domains and has made both ATE and BLI popular research topics over the past years.

The most successful strategies for finding cross-lingual equivalents rely on the distributional hypothesis or compositionality (see related research). The hypothesis of this project is that information from the monolingual term extraction phase (e.g., frequency, termhood and unithood statistics, part-of-speech (POS)) could be re-used as additional clues for finding equivalents cross-lingually. While it is not expected that these features alone will suffice to find cross-lingual equivalents, they might provide complementary information using features that have already been calculated for the monolingual ATE and could also help counter disadvantages of current approaches, such as the dependence on huge corpora. This pilot study was set up to test the potential informativeness of features from monolingual ATE to recognise term pairs cross-lingually in comparable corpora. First, an existing dataset for ATE (Rigouts Terryn et al., 2019b) was expanded with more cross-lingual annotations. Subsequently, a supervised binary classifier was constructed using only the features designed for monolingual term extraction. Further analyses of the classifier and the features illustrate how this information might complement the more established features.

2 Related Research

ATE from comparable corpora and BLI have received much research interest and certain trends have emerged. The distributional hypothesis appears to be the most popular approach for finding cross-lingual equivalents. This hypothesis states that equivalent lexical units will appear in similar contexts. The contexts of potential equivalents are compared by using some form of word vector representations. This can be done through the so-called standard approach (or a variation thereof).

In this case, context vectors are created for an n-word window around the lexical unit, subsequently, these vectors are normalised (e.g., using mutual information) and a bilingual seed lexicon is used to project between the source and target language. Once it is possible to map between the two vector spaces, a similarity measure (e.g., cosine similarity) can be calculated to measure the context similarity (Liu et al., 2018).

A second approach based on the distributional hypothesis consists in using neural networks to obtain word embeddings. An example is presented in (Hazem and Morin, 2018), where embeddings from the specialised corpus are combined with those from a larger, general corpus. Apart from word embeddings, they also experiment with character n-gram embeddings, which take into account the internal structure of words. This is another popular strategy, especially for morphologically-rich languages and the medical domain (Heyman et al., 2018; Hakami and Bollegala, 2017; Bollegala et al., 2015; Kontonatsios et al., 2013; Hazem and Morin, 2018). Character n-grams have repeatedly been shown to outperform word embeddings, or at least to be useful in combination with them. Since the previously described methods are mainly applied to single-word terms, there is another strategy specifically for multi-word terms, which is the compositionality approach, whereby each part of a multi-word term is translated separately to map it to potential equivalents. Such methods are highly reliant on bilingual dictionaries. Other common features are string similarity measures (Pinnis et al., 2019), Wikipedia-based features (Jakubina and Langlais, 2016, 2018), and temporal clues, burstiness and frequency (Irvine and Callison-Burch, 2013).

Some of the most commonly cited problems with current methodologies for ATE from comparable corpora are that they are dependent on very large resources for the context vectors, which is a big disadvantage for a task that has the specific goal of making bilingual lexicon building less reliant on expensive resources. While the increasing availability of large-scale, multilingual pre-trained language models (e.g., BERT (Devlin et al., 2019)) can be very helpful for BLI in general, it is less well-suited for multilingual ATE, since terminology is both less frequent and more domain-dependent than general language. Therefore, the specific characteristics of terms may not necessarily be captured well in these general language models, especially for those terms that also occur in general language but acquire a different meaning as a term in a specialised domain. A second, related disadvantage of most current approaches for ATE from comparable corpora is that they score badly for infrequent terms, even when ”infrequent” is broadly interpreted as having a frequency of up to 25 (Jakubina and Langlais, 2018). Other disadvantages are reliance on existing resources, such as bilingual seed lexicons or Wikipedia, separate methodologies for single- and multi-word terms (Liu et al., 2018) and, in the case of string similarity, dependence on similarities between source and target language. A final remark in this regard, is that it is very difficult to compare reported results. This is partly because of differences in methodology (e.g., entire ATE from comparable corpora pipeline or only classifying term pairs, focus on single- or multi-word terms, etc.) and evaluation measures (precision@rank, mean average precision and f1-score being the most common). Another important reason is the ambiguous nature of the task: determining whether two terms are equivalent is by no means straightforward. This can range from technical questions such as whether terms with an almost identical meaning, but from a different word class are considered correct, to more theoretical problems regarding the nature of equivalence (Le Serrec, 2012).

3 Data

For previous research on monolingual ATE (Rigouts Terryn et al., 2019b), three comparable corpora had been created in the domains of dressage, wind energy and heart failure, as well as one parallel corpus in the domain of corruption. All four corpora were constructed with English, French and Dutch texts. Around 50k tokens were manually annotated per domain/language, resulting in over 100k annotations of single- and multi-word terms and Named Entities (NEs). Cross-lingual annotations had already been added for the complete corpus on heart failure. A similar methodology was adopted to annotate cross-lingual equivalents in the other domains as well, although the annotations are less elaborate than those for the corpus on heart failure, which includes annotations of synonyms, abbreviations, alternative spellings, lemmas, hypernyms, hyponyms and other strongly related terms, in addition to cross-lingual equivalents. The annotations that were added for the other domains only concern cross-lingual equivalents. Moreover, the added annotations do not cover the entire corpora (as the original ones did), but they are sufficient to allow various cross-domain comparisons. The annotation work resulted in a total of over 3.5k validated term pairs per language pair (see Table 1). With the current methodology (see section 5), the order of the languages (English-French, English-Dutch, and French-Dutch) is irrelevant, so it could be reversed without influencing either the numbers or the results.

Domain	**EN-FR**	**EN-NL**	**FR-NL**
corruption	358	401	397
dressage	402	407	525
heart failure	2362	2467	2611
wind energy	425	598	389
Total	**3547**	**3873**	**3892**

Table 1: Number of positively validated term pairs per corpus

While the ultimate goal is to develop an entire pipeline for ATE from comparable corpora, from monolingual term extraction to bilingual term linking, the aim of the current pilot study was to test the potential of re-using information from the former for the latter. For this purpose, the previously mentioned annotations were transformed into datasets with positive (equivalent) and negative (non-equivalent) term pairs, which could be used as input for a binary classifier. All positive term pairs were manually annotated as valid equivalents and random sampling was used for negative examples, a methodology adopted in previous research as well (Kontonatsios et al., 2013; Hakami and Bollegala, 2017). For the sake of comparison, the methodology of Kontonatsios et al. (2013) was followed in other respects as well, for instance by starting with a balanced data set, with 50% positive and 50% negative instances. However, since this is not realistic in an actual pipeline for multilingual ATE from comparable corpora, imbalanced datasets were created as well with only 20% and 5% positive instances. The number of positives always remains the same (see Table 1), only the number of negatives varies according to these percentages.

A final note on the data is that the specialised corpora that were used are extremely small (50k tokens per language/domain) compared to the ones used in similar research (rarely under 1M tokens). Some of the features do refer to frequencies in large reference corpora (see section 4), but due to the specialised nature of the corpora and the fact that multi-word terms and NEs are included, many of the terms (single-word, multi-word and NE) have very low frequencies. For instance, out of the 3873 valid term pairs in the English-Dutch corpus, 1125 of the English source terms and 1340 Dutch target terms appear only once in the specialised corpus, and 1242 of the English terms and 2154 of the Dutch terms do not appear in any of the reference corpora. Considering that in similar research, terms appearing fewer than 25 times are considered to be infrequent (Jakubina and Langlais, 2016, 2018), it is interesting to see whether a decent performance can be obtained on such infrequent terms.

4 Monolingual ATE Features

The monolingual ATE features are based on the HAMLET tool (Rigouts Terryn et al., 2019a) and can be divided into 5 groups: *shape*, *frequency*, *statistics*, *related terms*, and *linguistics*. The number of features in each group and the description of these features can be found in Table 2. Most of these features have already been used for monolingual ATE, though most approaches are limited to a small number of these features. The reference corpora are Wikipedia dumps and news corpora in the respective languages, all limited to 10M tokens. For English, the News on Web corpus was used (Davies, 2017), for French the Gigaword corpus (Graff et al., 2011) and for Dutch the newspaper section of OpenSONAR (Oostdijk et al., 2013). The linguistic preprocessing was performed with the LeTs Preprocess toolkit (van de Kauter et al., 2013) and the part-of-speech (POS) tag sets of the three different languages were all mapped to a single set of 23 tags, so the same tags could be used across the languages. Preliminary experiments determined that the best way to encode the POS-patterns, was to have 3 vectors for all 23 individual tags: one for the tag of the first token of the term, one for the last and one for the frequency of all tags in the term. In the case of single-word terms, these would all be the same, but it was still an efficient way to encode the POS pattern for terms of varying lengths, without either losing too much

Feature group	#	Features
Shape	20	term length (in tokens or characters), capitalisation, presence of special characters
Frequency	12	relative frequency and document frequency of original term or lemmatised term in domain-specific corpus, newspaper reference corpus and Wikipedia corpus
Statistics	25	various termhood and unithood measures, calculated both for the original term and the lemmatised form (Vintar's termhood measure (Vintar, 2010)), C-value, TF-IDF, log-likelihood ratio, domain consensus, domain specificity, weirdness, basic, combo basic (more information about measures in (Astrakhantsev et al., 2015); measures that require a general reference corpus are calculated twice: once for the newspaper reference corpus, once for the Wikipedia reference corpus
Related Terms	12	count, combined frequency and average domain specificity of related terms, i.e. terms with the same lemma or normalised form and terms that are part of or contain the term in question
Linguistics	75	presence in stopword list, tag by automatic named entity recognition and POS, encoded as 3 one-hot vectors for the POS of the first token, POS of the last token and the frequency of all POS tags in the term

Table 2: Feature groups of the monolingual ATE with the number (#) of features in each group and a description of the features in that group

information or creating a disproportionate amount of POS-related features. There are no restrictions on term length, frequency or part-of-speech.

5 Experiments

5.1 Classifier and Features

By interpreting ATE from comparable corpora as a supervised binary classification task, we aim to test the usefulness of the monolingual ATE features for bilingual linking. Precision, Recall and f1-scores were calculated for each experiment. All experiments were performed with Python's scikit-learn package. Hyperparameter optimisation was performed through grid search and to counter the effect of random variations, the results of each experiment are averaged over 5 trials. Experiments were performed with either 5-fold cross-validation (within all domains of a single language pair or within one domain and language pair) or with a separate train and test set (test on one domain in one language pair and train on the three others). Preliminary experiments showed that the Random Forrest Classifier (RFC) and Multi-Layer Perceptron (MLP) outperformed the Decision Tree Classifier and the Logistic Regression Classifier. Since the RFC was more efficient than the MLP, had been used in previous research (Kontonatsios et al., 2013) and had a more stable performance, all further experiments were performed with the RFC. Positive instances (valid equivalents) were labelled as '1' and negatives (wrong equivalents) as '0'. The hyperparameter search space remained unchanged throughout the project ('min_samples_leaf': [5, 10], 'min_samples_split': [2, 10, 20], 'n_estimators': [150] and standard settings for all other hyperparameters), with the exception of 'class_weight', which varied from ['balanced', 0: 1, 1: 1.5, 0: 1, 1: 2, 0: 1, 1: 2.5] for the balanced dataset, to ['balanced', 0: 1, 1: 2, 0: 1, 1: 3, 0: 1, 1: 4, 0: 1, 1: 5, 0: 1, 1: 6] for the dataset with 20% positives and ['balanced', 0: 1, 1: 8, 0: 1, 1: 10, 0: 1, 1: 12, 0: 1, 1: 15] for the dataset with 5% positives.

As stated in section 4, the features are the ones used for monolingual ATE. There were two different setups to combine the features. In the first (CONCAT), the monolingual features of source and target term were simply concatenated, without any additional transformations. For the second (ABSDIF), the absolute difference was taken for all respective features. The features regarding the terms' POS pattern were analysed in more detail, since it was assumed that these features could potentially be very informative. Since there was no restriction on term length or POS pattern, the list of possible patterns across all languages is very long (200+ unique patterns). Therefore, as explained in the previous section, for the monolingual ATE, instead of a one-hot vector for all possible patterns, three (much shorter) vectors were used for all tags: one for the frequency of each tag in the pattern, one for the tag of the first token and one for that of the last token. While some information is lost this way, its compactness and ability to generalise was proven with good results for monolingual ATE. However, since POS pattern might be even more important for the bilingual linking, both approaches were tested and compared. Preliminary experiments showed better results (gain of 0.05 in f1-score) for the compact representations. Consequently, all further experiments were performed with this version of the features. Since only limited performance was expected from these features, it was decided to also test their compatibility with a string similarity feature (Levenshtein ratio), which seems intuitively more directly useful for the detection of equivalents in related languages, such as the ones used in this project. Using the python-Levenshtein package[1], Levenshtein ratio was calculated between all source and target terms. Before training the models, all features that showed no variance in the training data were removed. Generally, this affected some of the POS-features and special character features. Finally, the remaining features were scaled to [-1,1].

All these methodological difference lead to many different configurations: separate train/test sets versus 5-fold cross-validation, CONCAT versus ABSDIF features, with and without Levenshtein ratio, balanced dataset (50/50) versus slightly imbalanced dataset (80/20) versus very imbalanced dataset (95/5), and also three language pairs and four domains. Various experiments will be described in more detail in the following sections, but it can already be stated that the results were surprisingly good. The best obtained f1-score with Levenshtein features was 0.970 (precision 0.957 and recall 0.984). This was on a balanced dataset for the domain of corruption, French to Dutch, with ABSDIF features and separate train and test sets. The standard deviation of the f1-scores over the 5 trials was 0.002, indicating a rather stable performance. The best f1-score without Levenshtein features was still 0.939 (precision 0.911 and recall 0.970), with a standard deviation of 0.015. This was for the balanced data in the domain of dressage, French to Dutch, with CONCAT features and 5-fold cross-validation of only in-domain data. For comparison, the best reported state-of-the-art f1-score with a similar setup (supervised binary classifier with balanced data and 3-fold cross-validation) and of character n-grams features for English-French is 0.916 (Kontonatsios et al., 2013). Considering the nature of our features and the amount of infrequent terms in the data, our results compare much more favourably than expected and are a promising indication that features from monolingual ATE are relevant enough to be re-used for cross-lingual linking for ATE from comparable corpora.

5.2 Impact of Domain and Training Data

While domain can have a substantial effect on performance of monolingual ATE (Fedorenko et al., 2013; Conrado et al., 2013), performance across domains for our experiments with the cross-lingual linking of term equivalents appears to be largely domain-independent. For instance, f1-scores for experiments on the balanced datasets, using 5-fold cross-validation and averaged over experiments with different features are extremely similar: 0.928 (corruption), 0.927 (dressage), 0.928 (heart failure), and 0.930 (wind energy). Scores for more imbalanced datasets and with different features are comparably similar. This is somewhat surprising, considering that terms do have different characteristics in different domains (Rigouts Terryn et al., 2018, 2019b), that corruption is actually a parallel corpus and that there is much more data available for the domain of heart failure. The fact that corruption is a parallel, rather than a comparable corpus, should make it easier to find equivalents, but that fact may be compensated by the difficulty of the domain, since it was reported to be the most difficult to annotate (both monolingually and cross-lingually). Nevertheless, despite the similar results in this case, some of the highest obtained f1-scores were still obtained in the domain of corruption. As for the much larger size of the heart failure dataset: this may not affect the cross-validation experiments, but for the experiments with separate train and test sets, which use only training data from the other domains, heart failure does have a lower f1-score (averaged over all experiments with separate test set) than the other domains: 0.688 versus 0.811, 0.806, and

[1] https://pypi.org/project/python-Levenshtein/

0.800 in corruption, dressage, and wind energy respectively.

domain	p	r	f1
corruption	0.881	0.936	0.907
dressage	0.911	0.955	0.932
heart failure	0.875	0.953	0.912
wind energy	0.908	0.954	0.930
Average	**0.894**	**0.950**	**0.920**

Table 3: Precision (p), recall (r) and f1-scores (f1) per domain, averaged over all language pairs, on balanced datasets, without Levenshtein features, with concatenated features, using 5-fold cross-validation (and in-domain training data)

domain	p	r	f1
corruption	0.903	0.827	0.887
dressage	0.903	0.889	0.896
heart failure	0.829	0.868	0.848
wind energy	0.894	0.869	0.881
Average	**0.882**	**0.874**	**0.878**

Table 4: Precision (p), recall (r) and f1-scores (f1) per domain, averaged over all language pairs, on balanced datasets, without Levenshtein features, with concatenated features, using separate train and test sets (without in-domain training data)

While performance is stable across domains, training data does have an impact. Experiments with separate test sets (and only out-of-domain training data) perform worse than cross-validation experiments (with in-domain training data). Tables 3 and 4 show the results with the same experimental setup (balanced datasets, without Levenshtein features, with concatenated features) with cross-validation versus separate train and test sets. It is worth noting that, with different experimental configurations, the conclusions remain the same: with cross-validation, there is little to no difference in performance between domains, whereas separate train and test data results in slightly lower f1-scores for heart failure, the domain for which less training data is available. Moreover, performance is better for the former. In conclusion, while this methodology seems to work equally well for different domains, the presence of in-domain training data is important, and the amount of training data could also influence the scores.

lng.	**without Lev.**			**with Lev.**		
pair	**p**	**r**	**f1**	**p**	**r**	**f1**
en-fr	0.81	0.89	0.85	0.91	0.94	0.92
en-nl	0.78	0.90	0.83	0.91	0.93	0.92
fr-nl	0.79	0.91	0.85	0.94	0.96	0.95
Av.	**0.79**	**0.90**	**0.84**	**0.92**	**0.94**	**0.93**

Table 5: Precision (p), recall (r) and f1-scores (f1) per language pair, evaluated with 5-fold cross-validation on all domains of a language pair combined, evaluated on slightly imbalanced datasets (20% positives), with concatenated features, with and without Levenshtein features

5.3 Impact of Features and Language Pair

The impact of CONCAT versus ABSDIF features is minimal, with a slight advantage for CONCAT features (average difference in f1-score of 0.04). This is not surprising, since both contain almost the same information, and it indicates that the model is able to generalise well from concatenated features without any explicit link between equivalent features of source and target language terms. Still, a little information is lost by taking the absolute difference, so for future research it could be worth investigating other ways of combining the features. Since CONCAT features work best, the following experiments will all use these, unless stated otherwise.

The Levenshtein feature does have a large impact, as expected. Table 5 compares the results of two experiments with the same settings, with and without Levenshtein ratio as a feature. Since the difference in performance is more pronounced for imbalanced datasets (though it is noticeable as well on the balanced data), the reported results are for the dataset with only 20% positive instances. As can be seen, the models that include Levenshtein features achieve higher f1-scores, more specifically by increasing precision. This is true for all language pairs and also holds with other experimental settings. The only notable difference in this regard is between language pairs: including the Levenshtein feature has a bigger impact on the French-Dutch language pairs than on the others, which is somewhat unexpected, since the other language pairs seem more related (historically, English and French have influenced each other a lot and English and Dutch are both Germanic languages). No immediate explanation has been found to explain this phenomenon, especially since it is present in all domains and almost all configurations of the experiment. It is also reflected in the feature importance of Levenshtein ratio (see section 5.5). Apart from the Levenshtein feature, results for all language pairs are comparable for all settings.

5.4 Data Balance

As has already become clear, performance with balanced data is surprisingly good. However, in an actual pipeline for multilingual ATE from comparable corpora, this is not realistic, so the stability of the performance for imbalanced data was tested as well. Table 6 reports precision, recall and f1-scores for balanced (50/50), slightly imbalanced (80/20) and very imbalanced (95/5) data, using cross-validation to test on all domains of a language pair combined, and without Levenshtein feature. It should be noted that these scores are even higher when including Levenshtein ratio (f1-score for highly imbalanced data with Levenshtein is on average 0.902 with these settings).

Balanced data (50/50)			
	Precision	**Recall**	**f1-score**
en-fr	0.882	0.951	0.915
en-nl	0.885	0.953	0.917
fr-nl	0.889	0.957	0.921
Imbalanced data A (80/20)			
	Precision	**Recall**	**f1-score**
en-fr	0.810	0.888	0.847
en-nl	0.776	0.897	0.832
fr-nl	0.795	0.907	0.847
Imbalanced data B (95/5)			
	Precision	**Recall**	**f1-score**
en-fr	0.796	0.806	0.801
en-nl	0.755	0.813	0.783
fr-nl	0.753	0.741	0.794

Table 6: Precision (p), recall (r) and f1-scores (f1) per language pair evaluated with 5-fold cross-validation on all domains of a language pair combined, without Levenshtein features and for three differently balanced datasets

The first thing that can be seen in these tables, is that performance remains relatively high, despite the imbalance in the datasets. This will, of course, be partly due to the RFC's 'class_weight' parameter, but it is still promising, especially given the nature of the features. In all cases, recall is favoured over precision, even though precision never drops below 0.741. Conclusions are similar for all domains and with different experimental setups.

5.5 Feature Importance

To analyse the importance the model attributed to the various features, we looked at the models for the balanced dataset, created with 5-fold cross-validation on all domains combined per language pair. Conclusions across the language pairs are very similar, except that, when included, the Levenshtein feature gets a much higher importance for the French-Dutch language pair. Naturally, this feature is important in all models, but even more so for this language combination. For instance, in the models with ABSDIF features, the Levenshtein gets an importance of 20.8% for English-French, 24.5% for English-Dutch and 30.9% for French-Dutch. As mentioned in section 5.3, we have not yet been able to explain this difference satisfactorily. Since the models for all language pairs are similar in all other respects, the rest of the discussion will focus on a single language pair (English-French) as an example.

group	**feature**	**imp.**
SIM	Levenshtein	21%
STAT	Combo Basic	3.6%
LING	freq. of determiner POS tag	3.4%
SHAPE	nr. of tokens	2.9%
STAT	domain specificity of lemmatised form vs. Wikipedia	2.9%
STAT	domain specificity of original form vs. Wikipedia	2.7%
STAT	Vintar's termhood measure of original form vs. newspaper corpus	2.6%
LING	freq. of preposition POS tag	2.4%
LING	preposition as first POS tag	2.3%
SHAPE	nr. of characters	2.2%
REL	average domain specificity of all terms that contain the current term	2.1%

Table 7: Top ranked features with their feature groups and their attributed importance for the balanced en-fr models, created with 5-fold cross-validation on all domains combined, including Levenshtein features, with ABSDIF features

Table 7 shows all features that were attributed an importance of over 2% in a model with ABSDIF features. These results are for a model with Levenshtein features, but the ranking of the features remains similar without this feature. As can be seen, features from almost all feature groups are included (see also section 4): the string similarity feature (Levenshtein) (SIM), statistical (STAT) features, linguistic (LING) features, morphological/shape features (SHAPE), and related terms features (REL). The highest ranked frequency feature is not far behind in the ranking, in 18th place: relative frequency of the lemmatised form in the Wikipedia corpus (1.4%). Logically, features that show the least variance are also least important (e.g., features about rare special characters or rare first/last POS tags). Still, many features from many different groups are used.

Results with CONCAT features are more difficult to interpret, because the features from source and target term are separate. When included, Levenshtein ratio remains most important, but the other results differ. Strangely, the highest ranked features are all about the target language term; the first source term feature is only ranked 25th. Another difference with the ABSDIF models, is that, apart from the Levenshtein feature, the 15 highest ranked features are all statistical (12) or about related terms (3).

5.6 Error Analysis

To get a more in-depth idea of the performance, a limited error analysis was performed on one of the models. The results of an RFC model were analysed in English-Dutch, tested on the domain of dressage and trained on all other domains in the same language pair. This experiment used CONCAT features, including Levenshtein ratio and was performed on a balanced dataset. The f1-score for this particular run was 0.952 (precision 0.932 and recall 0.973). Out of 814 instances, there were 396 true positives, 378 true negatives, 11 false negatives and 29 false positives. Out of 11 false negatives, 4 contained numbers in either source or target language, which were written in full in the other language (e.g., *three-loop serpentine* and *slangenvolte met 3 bogen*). If the model has learnt to look at the presence of a number (shape feature), it is not surprising that equivalents where only one term contains a number are wrongly classified, even though a few other examples were correctly recognised despite this difficulty. Of the others, 5 concern either a source or target term that can be interpreted differently depending on the POS-tag, so the term pair may only be truly equivalent in some contexts (e.g., the English term *hoofs* and its Dutch equivalent *hoeven*, which can mean either *hoofs*, but also, more commonly, *ought to*). The remaining two concern pairs with no string similarity, and also different length: *equestrianism* and *equitation* as equivalents for *hippische sport* (some discussion is possible about the exact equivalence in this case). The false positives can be similarly explained. Only two are due to a coincidentally high Levenshtein ratio. Among the true positives, it is clear that even formally very different term pairs (e.g., *half-pass* and *appuyement*, or *inside hind leg* and *binnenachterbeen*) and infrequent terms can be correctly recognised with this methodology.

6 Conclusions and Future Research

The goal of this pilot study was to investigate whether features used for monolingual ATE could also be used to detect cross-lingual equivalents in comparable corpora. For this purpose, an existing dataset was expanded and these data were used to build binary classifiers in various experiments, testing the impact of certain features, domains, language pairs and the distribution of the dataset. Considering the models use none of the traditional features for this task and that the corpora were small and, therefore, contained many infrequent terms, the results were very promising and even outperformed some of the state-of-the-art approaches. Future research will have to determine whether these conclusions hold up in a complete pipeline for multilingual ATE from comparable corpora and whether and how they can best be combined with more typical features, e.g., distributional linking.

7 Acknowledgements

This research has been carried out as part of a PhD fellowship on the EXTRACT project, funded by the Research Foundation – Flanders.

References

Nikita Astrakhantsev, D. Fedorenko, and D. Yu. Turdakov. 2015. Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software* 41(6):336–349. https://doi.org/10.1134/S036176881506002X.

Danushka Bollegala, Georgios Kontonatsios, and Sophia Ananiadou. 2015. A Cross-Lingual Similarity Measure for Detecting Biomedical Term Translations. *PLOS ONE* 10(6). https://doi.org/10.1371/journal.pone.0126196.

Merley da Silva Conrado, Thiago A. Salgueiro Pardo, and Solange Oliveira Rezende. 2013. A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set. In *Proceedings of the NAACL HLT 2013 Student Research Workshop*. ACL, Atlanta, GA, USA, pages 16–23.

Mark Davies. 2017. The new 4.3 billion word NOW corpus, with 4–5 million words of data added every day. In *Proceedings of the 9th International Corpus Linguistics Conference. Birmingham*. Birmingham, UK.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* ArXiv: 1810.04805. http://arxiv.org/abs/1810.04805.

Denis Fedorenko, Nikita Astrakhantsev, and Denis Turdakov. 2013. Automatic recognition of domain-specific terms: an experimental evaluation. In *Proceedings of the Ninth Spring Researcher's Colloquium on Database and Information Systems*. Kazan, Russia, volume 26, pages 15–23.

David Graff, Ângelo Mendonça, and Denise DiPersio. 2011. French Gigaword Third Edition LDC2011t10. Technical report, Linguistic Data Consortium, Philadelphia, USA.

H. Hakami and D. Bollegala. 2017. A classification approach for detecting cross-lingual biomedical term translations. *Natural Language Engineering* 23(1):31–51. https://doi.org/10.1017/S1351324915000431.

Amir Hazem and Emmanuel Morin. 2018. Leveraging Meta-Embeddings for Bilingual Lexicon Extraction from Specialized Comparable Corpora. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, pages 937–949.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2018. A deep learning approach to bilingual lexicon induction in the biomedical domain. *BMC Bioinformatics* 19:259. https://doi.org/10.1186/s12859-018-2245-8.

Ann Irvine and Chris Callison-Burch. 2013. Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals. In *Proceedings of NAACL-HLT*. ACL, Atlanta, GA, USA, pages 518–523.

Laurent Jakubina and Philippe Langlais. 2016. A Comparison of Methods for Identifying the Translation of Words in a Comparable Corpus: Recipes and Limits. *Computación y Sistemas* 20(3):449–458. https://doi.org/10.13053/cys-20-3-2465.

Laurent Jakubina and Philippe Langlais. 2018. Reranking Candidate Lists for Improved Lexical Induction. In Ebrahim Bagheri and Jackie C.K. Cheung, editors, *Advances in Artificial Intelligence. Canadian AI 2018*, Springer International Publishing, Cham, volume 10832 of *Lecture Notes in Computer Science*, pages 121–132. https://doi.org/10.1007/978-3-319-89656-4$_1$0.

Georgios Kontonatsios, Ioannis Korkontzelos, Sophia Ananiadou, and Junichi Tsujii. 2013. Using a Random Forest Classifier to recognise translations of biomedical terms across languages. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Sofia, Bulgaria, pages 95–104.

Anaïch Le Serrec. 2012. *Analyse comparative de l'équivalence terminologique en corpus parallèle et en corpus comparable : application au domaine du changement climatique*. Doctor of Philosophy, Université de Montréal.

Jingshu Liu, Emmanuel Morin, and Peña Saldarriaga. 2018. Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, pages 2855–2866.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 219–247. https://doi.org/10.1007/978-3-642-30910-6$_1$3.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatjana Gornostaja, Špela Vintar, and Darja Fišer. 2019. Extracting Data from Comparable Corpora. In Inguna Skadiņa, Robert Gaizauskas, Bogdan Babych, Nikola Ljubešić, Dan Tufiş, and Andrejs Vasiļjevs, editors, *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*, Springer International Publishing, Cham, pages 89–139. https://doi.org/10.1007/978-3-319-99004-0$_4$.

Ayla Rigouts Terryn, Patrick Drouin, Véronique Hoste, and Els Lefever. 2019a. Analysing the Impact of Supervised Machine Learning on Automatic Term Extraction: HAMLET vs TermoStat. In *Proceedings of RANLP 2019*. Varna, Bulgaria.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2018. A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents. In *Proceedings of LREC 2018*. ELRA, Miyazaki, Japan.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2019b. In No Uncertain Terms: A

Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation* pages 1–34. https://doi.org/https://doi.org/10.1007/s10579-019-09453-9.

Marian van de Kauter, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste. 2013. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal* 3:103–120.

Spela Vintar. 2010. Bilingual Term Recognition Revisited. *Terminology* 16(2):141–158.

Extracting Bilingual Persian Italian Lexicon from Comparable Corpora Using Different Seed Dictionaries

Ebrahim Ansari,[†‡] **M.H. Sadreddini,**[§] **Mahsa Radinmehr,**[†] **and Ziba Khosravan**[†]
[†] Department of Computer Science and Information Technology,
Institute for Advanced Studies in Basic Sciences
[‡] Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University
[§] Computer Science and Engineering Department, Shiraz University
{ansari,radinmehr,zibakh}@iasbs.ac.ir
sadredin@shirazu.ac.ir

Abstract

Bilingual dictionaries are very important in various fields of natural language processing. In recent years, research on extracting new bilingual lexicons from non-parallel (comparable) corpora have been proposed. Almost all use a small existing dictionary or other resources to make an initial list called the "seed dictionary". In this paper, we discuss the use of different types of dictionaries as the initial starting list for creating a bilingual Persian-Italian lexicon from a comparable corpus. Our experiments apply state-of-the-art techniques on three different seed dictionaries; an existing dictionary, a dictionary created with pivot-based schema, and a dictionary extracted from a small Persian-Italian parallel text. The interesting challenge of our approach is to find a way to combine different dictionaries together in order to produce a better and more accurate lexicon. We propose two different novel combination models and examine the effect of them on various comparable corpora that have differing degrees of comparability. We conclude our work with a new weighting schema to improve the extracted lexicon. The experimental results show the efficiency of our proposed models.

1 Introduction

Bilingual lexicons are a key resource in a multilingual society. The availability of translation resources varies depending on the languages pairs. Therefore, bilingual dictionaries for languages with fewer native speakers are scarce or even nonexistent. Though automatic lexicon creation methods often have drawbacks such as including noise in the form of erroneous translations of some words, they are still popular because the alternative – manually constructing a dictionary – is very time-consuming. Automatic methods are often used to generate a first noisy dictionary that can be cleaned up and extended by manual work (Sjbergh, 2005).

A pivot language (bridge language) is useful for creating bilingual resources such as bilingual dictionaries. The Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries that share a common language. For example, using the Persian-English and the English-Italian dictionaries to build a new Persian-Italian lexicon. In recent years, some approaches based on this idea have been proposed (Tanaka and Umemura, 1994; Sjbergh, 2005; Istvn and Shoichi, 2009; Tsunakawa et al., 2008, 2013; Ahn and Frampton, 2006). In the last decade, some research has been proposed to acquire bilingual lexicons from non-parallel (comparable) corpora. These methods are based on this assumption: there is a correlation between co-occurrence patterns in different languages (Rapp, 1995). For example, if the words teacher and school co-occur more often than expected by chance in an English corpus then the German translations of teacher and school, *Lehrer* and *schule*, should also co-occur more often than expected in a German corpus (Rapp, 1995). Most of the approaches share a standard strategy based on context similarity. The basis of these methods is finding the target words that have the most similar distributions with a given source word. The starting point of this strategy is a list of bilingual expressions that are used to build the context vectors of all words in both languages. This starting list, or initial dictionary, is named the seed dictionary (Fung, 1995) and is usually provided by an external bilingual dictionary (Rapp, 1999; Chiao and Zweigenbaum, 2002; Fung and McKeown, 1997; Fung and Yee, 1998). Some of the recent methods use small parallel corpora to create their seed list (Otero, 2007) and some other use no dictionary for starting phases (Rapp and Zock, 2010). Sometimes there are different types of dictionaries, with each having its own accuracy. (Ansari et al., 2014) propose two simple methods to combine four different dictionaries (one existing dictionary and three dictionaries extracted using pivot based method) to increase the accuracy of the output. They use three languages English, Arabic, and French to create their pivot based lexicons. In this work, we use three different types of dictionaries and then combine them to create our seed dictionaries. The first dictionary is a small existing Persian-Italian dictionary. The second dictionary is extracted from a pivot-based method. The third dictionary is created from our small parallel Persian-Italian corpus. Using these dictionaries, we propose different combination strategies and a new weighting method to use on these different dictionaries.

2 Related works

In this Section, we discuss approaches and implementations in three parts and show their relation to our work.

2.1 Using Pivot languages

Over the past thirty years different approaches have been proposed to build a new source-pivot lexicon using a pivot language and consequently source-pivot and pivot-target dictionaries (Tanaka and Umemura, 1994; Istvn and Shoichi, 2009; Tsunakawa et al., 2008, 2013; Ahn and Frampton, 2006). One of the most known and highly cited methods is the approach of Tanaka and Umemura (Tanaka and Umemura, 1994) where they only use dictionaries to translate into and from a pivot language in order to generate a new dictionary. These pivot-language-based methods rely on the idea that the lookup of a word in an uncommon language through a third intermediated language could be done with machines. Tanaka and Umemura use bidirectional source-pivot and pivot-target dictionaries (harmonized dictionaries). Correct translation pairs are selected by means of inverse consultation. This method relies on counting the number of pivot language definitions of the source word, which identifies the target language definition (Tanaka and Umemura, 1994). Sjobergh presented another well-known method in this field (Sjbergh, 2005). He generated his English pivoted Swedish-Japanese dictionary where each Japanese-to-English description is compared with all Swedish-to-English descriptions. The scoring metric is based on word overlaps, weighted with inverse document frequency and consequently, the best matches are selected as translation pairs.

2.2 Using Parallel Corpora

Another way to create a bilingual dictionary is to use parallel corpora. Using parallel corpora to find a word translation (i.e. word alignment) started with primitive methods of (Brown et al., 1990) and continued with some other word alignment approaches such as (Gale and Church, 1991, 1993; Melamed, 1997; Ahrenberg et al., 1998; Tiedemann, 1998; Och et al., 1999). These approaches share a basic strategy of first having two parallel texts aligned in pair segments and second having word co-occurrences calculated based on that alignment. This approach usually reaches high score values of 90% precision with 90% recall, (Otero, 2007). Many studies show that for well-formed parallel corpora high accuracy rates of up to 99% can be achieved for both sentence and word alignment. Currently, almost the entire task of bilingual dictionary creation and especially the creation of a probability table for any word pairs could be done with well-known statistical machine translation software, GIZA++ (Och and Ney, 2003). Using Parallel corpora as the input of the dictionary creation process is attractive in two ways. First, alignment between sentences and words is very accurate as a natural characteristic of parallel corpora and these methods do not need any other external knowledge to build a bilingual lexicon. Second, no external bilingual dictionary (seed dictionary) is required. The main problem of creating a parallel corpus lexicon is the lack of extensive language pairs, therefore reliance on just using parallel corpora to build accurate bilingual dictionaries is impossible.

2.3 Using Comparable Corpora

There is a growing interest in the number of approaches focused on extracting word translations from comparable corpora (Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Djean et al., 2002; Kaji, 2005; Otero, 2007; Otero and Campos, 2010; Rapp and Zock, 2010; Bouamor et al., 2013; Irimia, 2012; E. Morin and Prochasson, 2013; Emmanuel and Hazem, 2014). Most of these approaches share a standard strategy based on context similarity. All of them are based on an assumption that there is a correlation between co-occurrence patterns in different languages (Rapp, 1995). For example, if the words teacher and school co-occur more often than expected by chance in a corpus of English, then the Italian translations of them, insegnante [teacher] and scuola [school] should also co-occur in a corpus of Italian more than expected by chance. The general strategy extracting bilingual lexicon from the comparable corpus could be described as follows:

Word target t is a candidate translation of word source s if the words with which word t co-occur within a particular window in the target corpus are translations of the words with which word s co-occurs within the same window in the source corpus.

The goal is to find the target words having the most similar distributions with a given source word. The starting point of this strategy is a list of bilingual expressions that are used to build the context vectors of all words in both languages. This starting list is called the seed dictionary. The seed dictionary is usually provided by an external bilingual dictionary. (Djean et al., 2002) uses one multilingual thesaurus as the starting list instead of using a bilingual dictionary. In (Otero, 2007) the starting list is provided by bilingual correlations previously extracted from a parallel corpus. In (Rapp et al., 2012), the authors extract a bilingual lexicon without using an existing starting list. Although they use no seed dictionary, their results are acceptable. Another interesting issue considered in recent years evaluating the effect of the degree of comparability on the accuracy of extracted resources (Li and Gaussier, 2010; Sharoff, 2013)

As described before, it is assumed that there is a small bilingual dictionary available at the beginning. Most methods use an existing dictionary (Rapp, 1999; Chiao and Zweigenbaum, 2002; Fung and McKeown, 1997; Fung and Yee, 1998) or build one with some small parallel resources (Otero, 2007). Entries in the dictionary are used as an initial list of seed words. Texts in both source and target languages are lemmatized and part-of-speech (POS) tagged with function words are removed. A fixed window size is chosen and it is determined how often a pair of words occurs within that text window. These windows are called the “fixed-size window” and word order does not take into account within a window. R. Rapp observed that word order of content words is often similar between languages, even between unrelated languages such as English and Chinese (Rapp, 1996). In approaches considering word order, for each lemma, there is a context vector whose dimensions are the same as the starting dictionary but in different window positions with regard to that lemma. For instance, if the window size is 2, the first context vector of lemma A, where each entry belongs to a unique seed word, shows the number of co-occurrences two positions to the left of A for that seed word. Three other vectors should also be computed, counting co-occurrences between A and the seed words appearing one position to the left of A and the same for two right hand positions following lemma A. Finally, all four vectors of length n are combined (where n is the size of the seed lexicon) into a single vector of length $4n$. This method takes into consideration the word orders to define contexts. In this paper, the efficiency of considering the word order schema is evaluated. Moreover, In the computation of the log-likelihood ratio, the simplified formula from Dunning and Rapp (Dunning, 1993) is used:

$$loglike(A,B) = \sum_{i,j \in 1,2} K_{ij} * \log \frac{K_{ij} * N}{C_i * R_j} \quad (1)$$

Therefore:

$$\begin{aligned} &loglike(A,B) = \\ &\quad K_{11} \log \frac{K_{11} * N}{C_1 * R_1} + K_{12} \log \frac{K_{12} * N}{C_1 * R_2} + \\ &\quad K_{21} \log \frac{K_{21} * N}{C_2 * R_1} + K_{22} \log \frac{K_{22} * N}{C_2 * R_2} \quad (2) \end{aligned}$$

Where:

$$C_1 = K_{11} + K_{12} \quad (3)$$

$$C_2 = K_{21} + K_{22} \quad (4)$$

$$R_1 = K_{11} + K_{21} \quad (5)$$

$$R_2 = K_{12} + K_{22} \quad (6)$$

$$N = C_1 + C_2 + R_1 + R_2 \quad (7)$$

With parameters K_{ij} expressed in terms of corpus frequencies:

K_{11} = frequency of common occurrence of word A and word *B*

K_{12} = corpus frequency of word A - K_{11}

K_{21} = corpus frequency of word B - K_{11}

K_{22} = size of corpus (no. of tokens) - corpus frequency of word *A* - corpus frequency of word *B*

For any word in a source language, the most similar word in a target language should be found. First, using a seed dictionary all known words in the co-occurrence vector are translated to the target language. Then, With consideration of the result vector, similarity computation is performed to all vectors in the co-occurrence matrix of the target language. Finally, for each primary vector in the source language matrix, the similarity values are computed and the target words are ranked according to these values. It is expected that the best translation will be ranked first in the sorted list (Rapp, 1999). Different similarity scores have been used in the variants of the classical approach (Rapp, 1999). In (Laroche and Langlais, 2010) the authors presented some experiments for different parameters like context, association measure, similarity measure, and seed lexicon. Some of the famous similarity metrics are included in the Appendix of this paper. We decided to use `diceMin` similarity score in our work which has been used previously in (Curran and Moens, 2002; Plas and Bouma, 2005; Otero, 2007). The `diceMin` score is the similarity of two vectors, X and Y, and is computed using the below similarity measure.

$$diceMin(X,Y) = \frac{2 \cdot \sum_{i=1}^{n} min(X_i, Y_i)}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i} \quad (8)$$

3 Our Approach

Our experiments to build a Persian-Italian lexicon are based on the comparable corpora window approach discussed in Section 2.3. An interesting challenge in our work is to combine different dictionaries with varying accuracies and use all of them as the seed dictionary for comparable corpora-based lexicon generation. We address this problem using different strategies: First, combining dictionaries with some simple priority rules, and then, using all translations together with and without considering the differences in their weights.

3.1 Building Seed Dictionaries

We have used three different dictionaries and their combinations as the seed dictionaries. The first dictionary is a small Persian-Italian dictionary named `DicEx`. For each entry, only the first translation is selected to create lemmas. While `DicEx` is a manually created dictionary, it is the most accurate dictionary in our experiments, and its size is the smallest in comparison with the other dictionaries. The second dictionary is created based on the pivot-based method presented in (Sjbergh, 2005), which contains top entries with the highest score. In contrast to the Sjobergh's implementation where the main focus is creating a dictionary with very large coverage, our goal is creating a small dictionary with more accuracy for use as a seed dictionary in the main system. Therefore, we select the top 40,000 translations from all translations and named it `DicPi`. Finally, the third dictionary is built using two little parallel Persian-Italian corpora which is named `DicPa`. When there is more than one translation for an entry in the primary dictionary, we should select one translation. Most standard approaches select the first translation in the existing dictionary or the candidate with the highest score in the extracted (created) dictionary. However, in (Irimia, 2012), several definitions for one word based on their scores could be selected in the seed dictionary generation step. Like other standard methods, we selected the first translation among all the candidates.

3.2 Using seed dictionaries to extract lexicon from Comparable Corpora

Mathematics and theoretical points of our approach were discussed in Section 2.3. Given that there are large differences between Persian and Italian words in syntax and grammar, the window-based approach is preferred. The baseline of the method implemented in our study is an adaptation of (Rapp, 1999). Based on our proposed idea, the seed dictionary could be an existing dictionary, an automatically created dictionary, or a combination of them. Previous approaches show the need for replacing the co-occurrence frequency in the matrix by measures that are able to eliminate word-frequency effects and consequently to favor significant word pairs. Therefore we use the log-likelihood ratio (i.e. Formula 1 (Dunning, 1993)) in our approach described in Section 2.3. To see its effect, we also carried out our tests without this metric by using the simple frequency matrix. In this experiment, we use `diceMin` similarity score as the preferred score. In Section 3.5 of this paper, a new similarity score, `newdiceMin` is proposed by the authors to weight dictionaries when different seed dictionaries are combined together.

3.3 Using simple combination

In this section, the process of creating the bigger seed dictionary by using a simple combination rule is discussed. `DicEx` has the highest accuracy and the accuracy of `DicPi` is higher than the dictionary created from the parallel corpus (i.e. `DicPa`). Based on the accuracy of dictionaries, a priority order is defined to create the final seed dictionary:

$$\texttt{DicEx} > \texttt{DicPi} > \texttt{DicPa}$$

Our simple combination rule is:

> Suppose that the priority of Dic$_i$ is more than the priority of Dic$_j$; if a word w is in both Dic$_i$ and Dic$_j$, its translation is selected from Dic$_i$ (i.e. the dictionary with higher priority)

By applying the above priority rule, a new Persian-Italian dictionary with more than 65,000 unique entries is created. We name this newly created dictionary `DicCoSi`. Apparently, all the words in `DicEx` are included in `DicCoSi`. The experimental results show an improvement in the extracted lexicon when this new dictionary `DicCoSi` is used as the main system's seed dictionary in comparison with using our three simple dictionaries individually.

3.4 Using independent word combination

In our simple priority-based combination which is described in Section 3.3, there is an important issue that should be discussed. Given two words, where the first one appears in all three dictionaries and the second one just appears in one dictionary. In our simple approach, there is no difference between these words. Therefore, a new advanced combination method is proposed. Our advanced combination method is based on the assumption that one word in two different dictionaries should be considered independently as two different words. For example, if a word appears in both dictionaries Dic$_1$ and Dic$_2$, it may have two independent columns in our vector matrix (i.e. it has two different weights in the transferred vectors). Therefore, the new dictionary named `DicCoIn` is created where its size is equal to the sum of our three dictionary's sizes. In this new dictionary, if the word x occurs in two dictionaries, there are two different entries for it named x$_i$ and x$_j$ where i and j are the indicators of corresponding dictionaries.

3.5 New weighting method

There is another problem in our proposed advanced combination. Even though some dictionaries are more accurate than others, there is no difference in dealing with initial seed dictionaries. In order to ease this problem, a new weighting model for similarity scores is introduced. This new metric relies on two following aspects:

(1) We could change the effect of each seed dictionary in order to consider the higher weight for the more accurate dictionary. All weights could be tuned manually.

(2) If a word appears in two dictionaries, then it is not necessary to count it twice as a double-count would produce an unfair skew. We could consider its weight a little bit more than a normal occurrence weight and then divide it between different dictionaries.

If there are k different dictionaries in our proposed independent word-based combination, to calculate the similarity scores between bilingual lemmas we could use the proposed equation:

$$newdiceMin(X,Y) = \frac{2 \cdot \sum_{j=1}^{k} \sum_{X_i \in Dic_j} min(X_i, Y_i) \cdot w_j}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i} \tag{9}$$

where n is the size of the new combined dictionary and w$_j$ is the weight of dictionary j. In our experiments, the size of k is equal to three. The new weighting method is based on this assumption that the dictionary with higher accuracy should affect the extracted lexicon more. In our experiments, two different sets of w$_j$ are studied and the results are evaluated in Section 5.1.

4 Preparing The Inputs

As stated prior, two primary inputs are needed to perform comparable corpora-based lexicon generation: seed dictionary and comparable corpus/corpora. Three different seed dictionaries are used in our experiments. Table 1 shows some characteristics of three dictionaries.

To evaluate the result, a test dataset is needed. The evaluation of the test is performed by two annotators. The first evaluator is one of the authors, who is a native Persian speaker and fluent in Italian and the second one is a Persian native who teaches the Italian language. If both of the evaluators agree on a translation word, it is accepted as a true translation, otherwise, the translation is considered false. We selected 400 Persian objective test words from Nabid Persian-English dictionary [1]. The frequencies of all the selected words in our corpora (general corpus and specific domain corpus) were greater than 100.

4.1 Seed Dictionaries

Dictionary Name	Entries	Mutual words
DicEx	13,309	N/A
DicPi	40,000	6,954
DicPa	40,000	4,220

Table 1: Number of entries and mutual words with DicEx of dictionaries used in our Experiments

In our experiments, three different types of comparable corpora are gathered: The first one is a small set of Wikipedia[2] articles in Persian and Italian. In order to skip those articles which are famous and well described in one of our languages (e.g. an article about an Italian village) we selected those article pairs where the difference between their sizes is not more than 50%. After applying this criterion, 6,500 articles are selected in both languages: about 150,000 sentences for Persian and 176,000 sentences in Italian. Both groups of sentences were tokenized and lemmatized. The resulting corpus is called `WikiCorpus` in our studies. This corpus is the most comparable corpus among our corpora (The comparability degree is more than the rest). The second corpus is the international sport-related news gathered from different Persian and Italian news agencies. We used the ISNA[3] and the FARS [4] for the Persian part, and the news agency CORRIERE DELLA SERA[5] and the Gazzetta dello Sport[6] for the Italian part. The numbers of selected articles are about 12,000 and about 15,000 from Persian and Italian resources, respectively. We named this corpus SportCorpus. We combined SportCorpus and `WikiCorpus` and used them together in our experimental results. We call this new combined corpus `SpeCorpus` (Specific domain-based corpus). The third corpus is based on international news gathered from different Persian and Italian news agencies. The difference between this corpus and `SpeCorpus` is that the former was gathered from sport-related news and this one is gathered from general subjects. This is our biggest corpus but obviously has a very low comparability degree in comparison with `SpeCorpus`. The number of articles in the Persian version was about 108,000 and for the Italian version was about 140,000 articles. We used ISNA and FARS news agencies for Persian version and CORRIERE DELLA SERA as the Italian resource. We named this corpus `GenCorpus`.

5 Experimental Results

All experiments described in this paper were applied on two types of comparable corpora: (1) the combination of `WikiCorpus` and `SportsCorpus` which we named `SpeCorpus`. (2) `GenCorpus` as a big, general, and less comparable corpus. The characteristics of these corpora were discussed in Section 4. In our experiments and for each test, two different result sets are calculated. The Top-1 measure is the number of times when the test word's acceptable translation is ranked first, divided by the number of test words. The Top-10 measure is equal to the number of times a correct translation for a word appears in the top 10 translations in the resulting lexicon, divided by the number of test words.

In the first phase of our experiments, all three previously mentioned dictionaries are used individually as the seed lexicon. These are the preexisting dictionary (`DicEx`), the pivot base extracted dictionary (`DicPi`) and the parallel corpus-based dictionary (`DicPa`). Figures 1 summarizes the

[1] Nabid Dictionary, written by Hani Kaabi, Iran, 2002

[2] https://www.wikipedia.org/

[3] https://isna.ir

[4] https://www.farsnews.com

[5] https://www.repubblica.it/

[6] La Gazzetta dello Sport, Italian, http://www.gazzetta.it/

evaluation results using these three seed dictionaries with and without using word order on `SpeCorpus`, the corpus with higher comparability degree. Figure 2 shows that using corpus with higher comparability degree increases the accuracy in both Top-1 and Top-10 results significantly. As it is expected, this difference for Top-1 results is more than the Top-10 measure. According to the results, the `DicEx` has better outcome despite its small size compared with the other dictionaries. A reason is the high accuracy of `DicEx` as it is a handmade dictionary. We could consider it a 100% accurate dictionary. The experimental results show that `DicPi` has a slightly better efficiency in comparison with parallel corpora based dictionary `DicPa`. The authors conclude that the reason is the limitation of our parallel Persian-Italian corpus used to create the translation table.

In the second part of our experiments, we evaluated our ideas of combining different dictionaries together. Table 2 shows the results of this study. According to this table, the best results for Top-1 measure belong to the simple combination model when all dictionaries are combined together. The best Top-10 results belong to the advanced combination model combining all dictionaries. In advanced combination, all words in all dictionaries are selected in the lexicon generation phase, and this generally gives us better Top-10 results. An important issue for our advanced combination is that all translations in different dictionaries have the same weight and this may decrease the effect of `DicEx`. Although it is our most accurate dictionary, it is also the smallest one. This problem is tackled in the next section by using our weighting lemma.

5.1 Using new weighting

Two different heuristics are considered to adjust weights in our weighting schema. The first one is to tune weights based on dictionaries accuracy. The accuracies could be collected from Top-10 scores calculated in our experiments. In the first set, the weights for `DicEx`, `DicPi` and `DicPa` are 0.7, 0.64 and 0.59, respectively. In the second heuristic set, the weights are calculated based on both accuracy and the dictionary size. This weight set is constructed based on the assumption that the bigger dictionary should have a lower effect on the final result. We used the following formula to calculate the weights.

$$w_i = accuracy_i \cdot \frac{MaxSize}{size_i} \tag{10}$$

Based on the second heuristic, and with considering the results in our study the weights are:

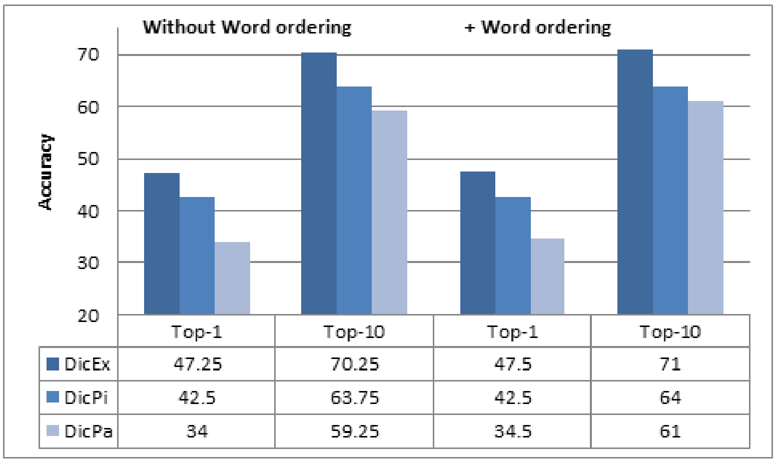
$$\mathrm{W}_{\texttt{DicEx}} = 2.10,$$
$$\mathrm{W}_{\texttt{DicPi}} = 0.64,$$
$$\mathrm{W}_{\texttt{DicPa}} = 0.59.$$

The results of these experiments based on different weighting sets are shown in Table 3. $\mathrm{W}_i = 1$ presents the classic approach without using the proposed weighting system.

Finally, Figure 3 shows a brief demonstration to see the effect of our combination methods in comparison with classic approaches when they used just the existing dictionary, `DicEx` (the most accurate independent dictionary in our study) as the seed dictionary. In all results, the log-likelihood ratio with considering word ordering schema are used to extract bilingual lexicons from `SpeCorpus`, our corpus with high comparability degree. `AC` stands for advanced combination model.

6 Conclusion

In the last decade, some approaches have been proposed to extract bilingual lexicons from comparable corpora. In order to create a Persian-Italian lexicon, we decided to implement a comparable corpora-based lexicon generation method. In our study, three different seed lexicons (and combinations) are used consisting of one pre-existing dictionary and two extracted dictionaries. The first extracted dictionary is based on parallel-corpora dictionary creation methods and the second one is extracted by pivot language models. While for a seed dictionary a small dictionary is needed, we just selected the top translations from these created dictionaries. In the first part of our study, the effects of using these dictionaries on different types of comparable corpora are evaluated. A new and interesting challenge which is introduced in this paper was creating a new seed by combining some different dictionaries. We used two different strategies: First, composing dictionaries with some priority rules; second, using all dictionaries together considering similar words in two dictionaries as a different word. Both of these strategies were studied and based on our experimental



	Top-1	Top-10	Top-1	Top-10
■ DicEx	47.25	70.25	47.5	71
■ DicPi	42.5	63.75	42.5	64
■ DicPa	34	59.25	34.5	61

Figure 1: Results of using independent dictionaries with and without considering word orders. All results are based on log-likelihood measurement using SpeCorpus (in-domain corpus)

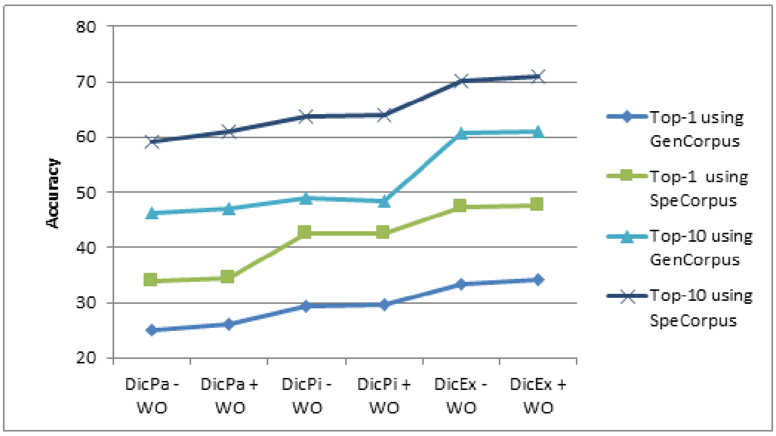


Figure 2: Effect of using different corpora in with different comparability degree

results these novel dictionary combinations could improve the efficiency of the results. Furthermore, the effect of comparability degree of the initial comparable corpus is studied using different types of comparable corpora. Finally, a new weighting method has been proposed to increase the efficiency of our dictionary combination. This new weighting method uses the assumption that the effect of a more accurate seed dictionary should have a better result in comparison with others.

Acknowledgments

The authors gratefully acknowledge the contribution and help of Dr. Fatemeh Alimardani, Dr. Daniele Sartiano, Vahid Pooya, Dr. Amir Onsori, S. M. H. Mirsadeghi, and Dr. Mahshid Nikravesh to this work. The research was partially supported by OP RDE project No. CZ.02.2.69/0.0/0.0/16_027/0008495, International Mobility of Researchers at Charles University.

References

Kisuh Ahn and Matthew Frampton. 2006. Automatic generation of translation dictionaries using intermediary languages. Association for Computational Linguistics, 1608848, pages 41–44.

Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. Association for Computational Linguistics, 980851, pages 29–35. https://doi.org/10.3115/980451.980851.

Ebrahim Ansari, M. H. Sadreddini, Alireza Tabebord-

Dictionary Name	Top-1		Top-10	
	Simple	Advanced	Simple	Advanced
DicEx + DicPi	50.00	49.50	75.00	75.50
DicEx + DicPa	48.75	48.00	74.00	74.75
DicPi + DicPa	42.50	43.00	66.75	67.50
All Dictionaries	**50.25**	49.75	75.25	**76.75**

Table 2: The effect of different dictionaries in combination with different methods on SPECORPUS for advanced combination

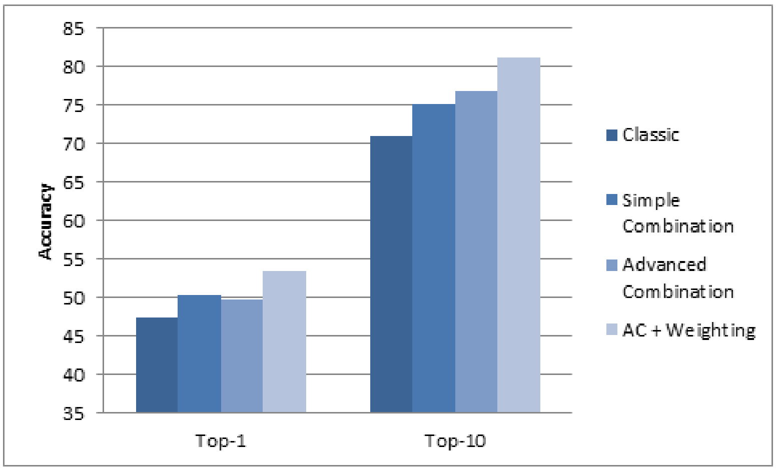


Figure 3: The effect of different introduced combinations

	Top-1	Top-10
W_i=1	50.25	76.75
Weight 1	52.50	78.25
Weight 2	**53.75**	**81.25**

Table 3: The effect of new weighting schema on accuracy of extracted dictionary (In all tests, the combination of three dictionaries is used and the comparable corpus is SPECORPUS)

bar, and Mehdi Sheikhalishahi. 2014. Combining different seed dictionaries to extract lexicon from comparable corpus. *Indian Journal of Science and Technology* 7(9):1279–1288.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Building specialized bilingual lexicons using word sense disambiguation. pages 952–956.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.* 16(2):79–85.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. Association for Computational Linguistics, 1071904, pages 1–5. https://doi.org/10.3115/1071884.1071904.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. Association for Computational Linguistics, 1118635, pages 59–66. https://doi.org/10.3115/1118627.1118635.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 19(1):61–74.

Herv Djean, ric Gaussier, and Fatia Sadat. 2002. Bilingual terminology extraction: an approach based on a multi-lingual thesaurus applicable to comparable corpora.

B. Daille E. Morin and E. Prochasson. 2013. Bilingual terminology mining from language for special purposes comparable corpora. In *Building and Using Comparable Corpora*. Springer.

Morin Emmanuel and Amir Hazem. 2014. Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 1284–1293.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. pages 173–183.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations fromNon-parallel corpora. volume 1, pages 192–202.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. Association for Computational Linguistics, 980916, volume 1, pages 414–420. https://doi.org/10.3115/980451.980916.

William A. Gale and Kenneth W. Church. 1991. Identifying word correspondence in parallel texts. Association for Computational Linguistics, 112428, pages 152–157. https://doi.org/10.3115/112405.112428.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comput. Linguist.* 19(1):75–102.

Elena Irimia. 2012. Experimenting with extracting lexical dictionaries from comparable corpora for: English-romanian language pair. pages 49–55.

Varga Istvn and Yokoyama Shoichi. 2009. Bilingual dictionary generation for low-resourced language pairs. Association for Computational Linguistics, 1699625, volume 2, pages 862–870.

Hiroyuki Kaji. 2005. Extracting translation equivalents from bilingual comparable corpora. *IEICE - Trans. Inf. Syst.* E88-D(2):313–323. https://doi.org/10.1093/ietisy/E88-D.2.313.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. Association for Computational Linguistics, 1873851, pages 617–625.

Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. Association for Computational Linguistics, 1873854, pages 644–652.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. Association for Computational Linguistics, 979656, pages 305–312. https://doi.org/10.3115/979617.979656.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.* 29(1):19–51. https://doi.org/10.1162/089120103321337421.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. pages 20–28.

Pablo Gamallo Otero. 2007. Learning bilingual lexicons from comparable english and spanish corpora. pages 191–198.

Pablo Gamallo Otero and Jose Ramom Pichel Campos. 2010. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. Springer-Verlag, 2175399, pages 473–483. https://doi.org/10.1007/978-3-642-12116-6_40.

Lonneke van der Plas and Gosse Bouma. 2005. Syntactic contexts for finding semantically similar words. pages 173–186.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. Association for Computational Linguistics, 981709, pages 320–322. https://doi.org/10.3115/981658.981709.

Reinhard Rapp. 1996. Die berechnung von assoziationen: ein korpuslinguistischer ansatz. *Hildesheim; Zrich; New York: Olms* .

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. Association for Computational Linguistics, 1034756, pages 519–526. https://doi.org/10.3115/1034678.1034756.

Reinhard Rapp, Serge Sharoff, and Bogdan Babych. 2012. Identifying word translations from comparable documents without a seed lexicon. pages 460–466.

Reinhard Rapp and Michael Zock. 2010. Utilizing citations of foreign words in corpus-based dictionary generation .

Serge Sharoff. 2013. Measuring the distance between comparable corpora between languages. In *BUCC: Building and Using Comparable Corpora*. Springer.

Jonas Sjbergh. 2005. Creating a free digital japanese-swedish lexicon. pages 296–300.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. Association for Computational Linguistics, 991937, pages 297–303. https://doi.org/10.3115/991886.991937.

Jrg Tiedemann. 1998. Extraction of translation equivalents from parallel corpora.

Takashi Tsunakawa, Naoaki Okazaki, and Junichi Tsujii. 2008. Building bilingual lexicons using lexical translation probabilities via pivot languages. pages 1664–1667.

Takashi Tsunakawa, Yosuke Yamamoto, and Hiroyuki Kaji. 2013. Improving calculation of contextual similarity for constructing a bilingual dictionary via a third language. pages 1056–1061.

From the cultivation of comparable corpora to harvesting from them: A quantitative and qualitative exploration

Benjamin K. Tsou
City University of Hong Kong
Hong Kong University of Science and Technology
Chilin (HK) Ltd
btsou99@gmail.com

Kapo Chow
Chilin (HK) Ltd
kapo.rclis@gmail.com

Abstract

This paper reports on a relatively new but important area involving: (1) The corpus cultivation of comparable multilingual patents for, (2) The building of a large-scale parallel sentence corpus, and thence, (3) The harvesting from them of useful language resources for NLP and other applications. Three major efforts are reported: (a) The sourcing and cultivation of a large corpus of bilingual and comparable patents suitable for diverse applications in strategic NLP, (b) The mining of high-quality bilingual aligned sentences from the comparable patents which contain many technical terms, and (c) The extraction of bilingual multi-word expressions (MWEs) from the parallel sentences in (b), where there is often no simple isomorphic cross-lingual correspondences among the lexical items. An analysis is provided on how over 1 million very high quality lexical entries consisting of bilingual multi-word expressions and their multiple renditions have been harvested in the initial and expanding version of a derived MWE database. This has followed a series of rigorous winnowing of an initial large database of 10 years of English and Chinese patents consisting of more than 5,000 million English words and 12,000 million Chinese characters. Some issues on efficacy in data curation to maximize both quantity and quality are raised, as well as an outline of how the MWEs with their multiple renditions are put to good use by translators and trainers of translators.

1 Introduction

To be useful for NLP, the size of corpora must be quite large. In recent decades, there have been interests in working with big and related corpora in different languages, where the degree of comparability may range widely from parallel texts to even unrelated texts (Sharoff et al., 2013). Such interests were boosted by, for example, the official production of bilingual corpora (Germann, 2001; Koehn, 2005). Common examples are voluminous bilingual parliamentary records and parallel bilingual legal codes, for example, from governments or international organizations, as well as the availability of multilingual comparable corpora from Wikipedia. The availability of other bilingual texts which are comparable, if not parallel, has boosted considerably advances in machine translation and other NLP efforts (Sharoff et al., 2013; Zhao and Vogel, 2002; Resnik and Smith, 2003; Munteanu and Marcu, 2005; Wu and Fung, 2005; Smith et al., 2010). However, comparable corpora as well as similar and useful monolingual corpora also exist in other relatively untapped domains and areas which are of strategic importance. This is especially the case with patents in the cross-lingual context which impinge on complex global trade and economic competition as their content involves advanced scientific and technological developments which require comprehensive but succinct description and where ownership of intellectual property rights may be contentious and have to be protected legally.

2 Corpus Cultivation: From Quantity to Quality

2.1 Stage A: From Big Data to Useful Data

There can be several stages in the long work flow to go from Big Data to useful data and, it is no simple task which can be accomplished semi-automatically. It is often assumed that the more data the better, and that the Age of Big Data would provide easy and theoretically limitless access to data. In the case of patents, they are documents which usually contain much useful information and new technical terms, and may be available in more than one language because inventions described in a patent may only be best protected in the country where it is filed. As a result, an applicant who wishes to protect his invention would file the patent in other countries in the local languages. Usually, the two or more versions should be parallel but there could be also textual variations to tactfully protect the legal rights of the technical content and because of non-uniformity in human efforts. There are also cross-references to other relevant patents, and quite often there are bilingually paired patents or bilingually paired textual segments within clusters of comparable patents.

At the start, we took 10 years of Chinese and English patents officially published around the turn of the millennium. The combined size of these English and Chinese patents is very large:

English: 5,351.7M words
Chinese: 12,001M characters

This enormous quantity is based on the number of English and Chinese patents published during this decade: English patents: 840,027 documents and Chinese patents: 967,686 documents, and based on the average size of more than 300,000 English and Chinese patents which have been analyzed. (Lu et al.,2016)

2.2 Stage B: Corpus Cultivation of Comparable patents

The identification and winnowing of comparable patents from stage A begins with the meta-information of the patents. Essentially the cross-references in the section "Worldwide applications" are examined.

From the official 2009 website of the State Intellectual Property Office (SIPO) in China, about 200K Chinese patents were found to have links with previously filed PCT applications in English and we crawled their bibliographical data, titles, abstracts and the major claim from the Web. Other claims and descriptions were also added in the processing. All PCT patent applications are filed through WIPO. Drawing on the Chinese patents mentioned above, the corresponding English patents were searched from the WIPO website to obtain relevant sections of the English PCT applications, including bibliographical data, title, abstract, claims and description. About 80% (160K) of the Chinese patents have corresponding English ones and a total of about 340K comparable bilingual patents form the initial base corpus of comparable patents. (Lu et al., 2015, 2016)

These cultivation efforts involved considerable manual efforts and have yielded the following combined size of textual content .

English: 1,020.4M words
Chinese: 1,986.4M characters

2.3 Stage C*1*: Sentence Alignment 1

Following stage B our preliminary efforts produced a drastically reduced set of bilingual English-Chinese parallel sentences by means of iterative bilingual sentence alignment.

First we use a bilingual seed dictionary to preliminarily align the sentences in each section (abstract, claims, descriptions) of the comparable patents, and perform filtering using length-based (Gale and Church, 1991) and dictionary-based scores. The dictionary-based similarity score P_d of a sentence pair is computed based on a bilingual dictionary as follows (Utiyama and Isahara, 2003):

$$p_d(S_c, S_e) = \frac{\sum_{w_c \in S_c} \sum_{w_e \in S_e} \frac{\gamma(w_c, w_e)}{\deg(w_c)\deg(w_e)}}{(l_e + l_c)/2}$$

where w_c and w_e are respectively the word types in Chinese sentence S_c and English sentence S_e; l_c and l_e respectively denote the lengths of S_c and S_e in terms of the number of words; and $\gamma(w_c, w_e)$ = 1 if w_c and w_e is a translation pair in the bilingual dictionary or are the same string, otherwise 0; and

$$\deg(w_c) = \sum_{w_e \in S_e} \gamma(w_c, w_e)$$

$$\deg(w_e) = \sum_{w_e \in S_c} \gamma(w_c, w_e)$$

For the bilingual dictionary, we drew from three publications: viz, LDC_CE_DIC2.0 (http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm), bilingual terms in HowNet (http://dict.cnki.net/) and the bilingual lexicon in Champollion (Ma, 2006). We then removed sentence pairs using length filtering and ratio filtering by two means: 1) For length filtering: if a sentence pair had more than n words in the English sentence or more than m characters in the Chinese one, it was removed; 2) For length ratio filtering: we discarded sentence pairs with

Chinese-English length ratio outside the range of 0.8 to 1.8. The parameters here were set empirically. We further filtered the parallel sentence candidates by learning an IBM Model-1 algorithm (Brown et al., 1993; Och and Ney, 2004) on the remaining aligned sentences and computing the translation similarity score P_t of sentence pairs by combining the translation probability value of both directions (i.e. Chinese->English and English->Chinese) based on the trained IBM-1 model (Moore, 2002; Chen, 2003; Lu et al, 2009). It is computed as follows:

$$p_t(S_c,S_e)=\frac{log(P(S_e\mid S_c))+log(P(S_c\mid S_e))}{l_c+l_e}$$

where $P(S_e\mid S_c)$ denotes the probability that a translator will produce S_e in English when presented with S_c in Chinese, and vice versa for $P(S_c\mid S_e)$. Sentence pairs with similarity score P_t lower than a predefined threshold are filtered out as incorrect aligned sentences.

After the end of stage C1, we have an initial bilingual sentences corpus with the following combined size.

English: 355.3M words
Chinese: 628.2M characters

2.4 Stage C2: Sentence Alignment 2 (with enriched bilingual glossary)

Cyclical sentence alignment is applied with the help of an enriched bilingual glossary. This has enhanced the size of bilingual sentence corpus as follows.

English: 989.2M words
Chinese: 1,914.6M characters

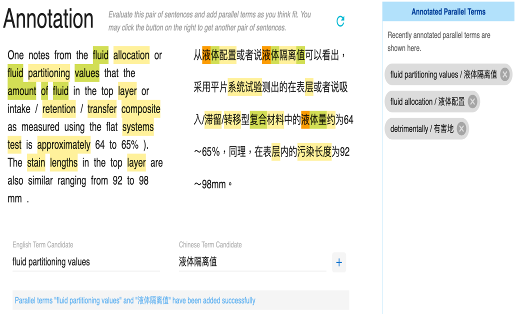
It is noteworthy that the results show near maximal recall as can be seen from the ceiling effect when compared with the results of stage B. At the same time, it reflects on the efficacy of the sentence alignment efforts in stage B.

2.5 Semi-automatic Curation Efforts

Whereas stage A identified patents whose contents are parallel, if not comparable, stage B identified linguistic segments within them which are comparable on objective and statistical basis.

From stage B, automatic multi-word expression extraction produced high-frequency bilingual terms. But within the parallel corpus there were residual and low frequency but valid useful terms for stylistic or other reasons. They could be further exploited. Because of their rarity and sparsity, there was no effective automatic means to extract these low-frequency bilingual terms. We thus developed a computer-assisted bilingual term extraction system to allow annotators to perform this task in a simple but effective manner. The system has been a web-based online system for annotators to mine unextracted term pairs in the following steps:

First, pairs of aligned sentences are randomly selected from the pool of about 40M candidate sentence pairs. For these pairs, all recognized bilingual terms from the current bilingual glossary are highlighted. To enhance visual presentation, words that are covered by multiple bilingual terms are highlighted in a darker shade. This allows annotators to focus on unextracted terms at a glance, and highlight them in both languages, thereby adding the selected term pair to the database. Once annotation of all unextracted bilingual terms are completed, the annotator can choose to view another random pair of sentences from the system, and repeat the same examination and annotation process.



Annotation

Evaluate this pair of sentences and add parallel terms as you think fit. You may click the button on the right to get another pair of sentences.

One notes from the fluid allocation or fluid partitioning values that the amount of fluid in the top layer or intake / retention / transfer composite as measured using the flat systems test is approximately 64 to 65%). The stain lengths in the top layer are also similar ranging from 92 to 98 mm .

从液体配置或者说液体隔离值可以看出，采用平片系统试验测出的在表层或者说吸入/滞留/转移型复合材料中的液体量约为64~65%，同理，在表层内的污染长度为92~98mm。

English Term Candidate: fluid partitioning values

Chinese Term Candidate: 液体隔离值

Parallel terms "fluid partitioning values" and "液体隔离值" have been added successfully

Annotated Parallel Terms

Recently annotated parallel terms are shown here.

- fluid partitioning values / 液体隔离值
- fluid allocation / 液体配置
- detrimentally / 有害地

From the sample user interface shown above, it can be seen that two pairs of terms, “fluid allocation” / “液体配置” and “fluid partitioning values” / “液体隔离值”, are newly marked and added to the database. The system underlying this curation process is outlined in the next section, which focuses on the systematic extraction of linguistically well-formed multi-word expressions.

2.6 Stage D*n*: Multi-word Expressions

We implemented an automatic procedure to acquire the bilingual phrases from the above parallel sentence pair corpus on the basis of Tian et al. (2011, 2014) and discussed in Tsou et al. (2017b). We first derived the phrases from monolingual data, i.e. the source sentences of parallel data, by considering the possible n-gram of words. For the high frequency words or phrases, we evaluated how likely a phrase could be constructed by two (words or) phrases p_i and p_j, by considering the following significant score:

$$\text{Score}(P_1, P_2) = \frac{f(P_1 \oplus P_2) - \mu_0(P_1 - P_2)}{\sqrt{f(P_1 \oplus P_2)}}$$

where $f(\cdot)$ and $\mu_0(\cdot)$ were the frequency and the mean under null hypothesis of independence of two phrases (El-Kishky et al., 2014). $\oplus$ was the concatenation operator. The equation computes the number of standard deviations away from the expected number of occurrences, and this score could be considered a generalization of the t-statistic for identifying dependent bigrams. To extend the identified phrases to its bilingual, we first derived the word alignment information for the bilingual data using the model proposed by Dyer et al. (2013). The word alignments acted as vital information and were used to project the phrase boundaries from the source sentences to the target side of the bilingual data (Zeng et al., 2014). For those unaligned words or phrases, we simply ignored them from the induction process. This bilingual phrases extraction model was based on an unsupervised approach, where all the statistics were automatically derived from a given parallel corpus aligned at sentence level.

The bilingual multi-word expressions obtained were fed into an online translation engine for further automatic evaluation. Each monolingual part of a pair of bilingual terms was input to a translation engine whose output was compared with the other part of the bilingual pair in terms of Levenshtein distance (LD) (Haldar et al. 2011). The results were as follow: LD=0:11.9%, LD=1:25.5%; LD=2:29.3%; LD=3:11.9%; LD=4: 11.0%; LD≥ 5:10.3%.

We noted that from the results, only 11.9% were identical with online translation results. The remaining nearly 90% were not identical but were still potentially valid entries. They were more valuable because they reflected the actual alternate language use in this particular domain which would be usually ignored by automatic means and also was not readily available through any public resources. Further empirical studies showed that lower edit-distance entries require less manual modifications. When the edit distance was 1 or 2, about 65% were valid entries without modifications and 5% more could be useful after manual modifications. For distance 3 or 4, about 55% were valid entries, while about 10% could become useful after modifications.

Our efforts so far have produced over 6 million MWE candidate entries. Furthermore, mostly straightforward manual checks have already yielded 1 million good bilingual MWEs and more items are expected after iterative processing. The human efforts have mostly involved the pruning of redundant constituents and in some cases the recovery of missing constituents.

We note there is a noticeable drastic reduction in going from the parallel aligned sentences and sentence fragments to bilingual terms, including MWE's. Following further filtering and human supervision are found in the following sub-corpus only linguistically well-formed expressions.

English: 2.95M words
Chinese: 5.89M characters

2.7 Pairing Bilingual Terms

Based on the bilingual MWE database, we have constructed a cross-lingual search system – *Chilin PatentLex*. The following are some examples of search results. Based on the meta information of each patent, we are able to provide insightful statistics through the search query, as can be seen in Table1.

2.7.1 "Channel" – Alternate Renditions in Chinese Patents

PatentLex (%)	PatentLex (%)	PatentLex (%)
道 (48.58) 信道 (30.89) 通道 (10.92) 频道 (3.06) 槽 (2.55) 沟道 (1.98) 腔 (0.71)	沟槽 (0.58) 管道 (0.37) 道宽 (0.12) 渠道 (0.06) 波道 (0.02) 途径 (0.02) 路线 (0.01)	水道 (0.01) 槽钢 (0.01) 沟渠 (0) 通道化 (0) 海峡 (0) 管箱 (0) 通风槽 (0) 窜槽 (0)

Table 1 Alternate Chinese Renditions for "*Channel*"

In Table 1, we note that the English term "channel" has 22 renditions in Chinese and that the percentage distribution of actual usage of each alternate rendition shows considerable variations. Among the 22 actual alternate renditions, 3 are used in 10% or more of the total, while the majority has frequency less than 1% (0% indicates very low usage of less than 0.01). While this is exhaustive for the 10 years of Chinese-English patent we collected, it may be noted that the different IPC domains have not been equally represented in patents. Given further cultivation of the comparable database and also breakdown according to IPC domains, we expect the distribution of the alternate renditions to change and be more balanced.

From the current database, we can see that the highest percentage of usage comes from the single-character word 1."道", followed by 2. "信道" and 3."通道". It is worth noting that since the percentage comes from figures gathered from string matching of the rendition with the parallel sentence pairs, the rendition “信道” will also contribute to the counting of “道”. This may lower the *precision* but can increase *recall* and lead to more relevant authentic example sentences for users to examine.

We can further retrieve all the example sentence pairs containing MWE’s of “channel” with various renditions from different IPC domains (see examples in Table 4), and even with low frequencies of usage, so that the needs of the translator may be met.

Furthermore, beyond the renditions of the search keyword “channel”, the search engine will also return other fuzzy results with MWEs containing “channel”, again with their respective renditions and distribution, as can be seen in Table 1 and Table 2.

English Matched Term	Chinese Renditions (%)
a corresponding channel	对应信道(100)
a plurality of channels	多个通道(100)
a second counting channel	第二计数通道(100)
absolute grant channel	绝对许可信道(100)

access channel	1. 接入信道(78.16)
	2. 访问信道(17.84)
	3.存取信道(3.38)
	4.进入通道(0.56)
	5.出入通道(0.04)

Table2 Fuzzy Search Results (source: PatentLex)

2.7.2 Cytidine

The fuzzy search results of the multi-word term “cytidine” from two sources are provided for comparison below: PatentLex and HOWNET, a well-known Chinese language resource.

Cytidine
胞苷(91) 胞嘧啶核苷(16)
(source: HOWNET)

English Matched Term	Chinese Renditions (%)
cytidine	胞苷(100)

5-azacytidine	氮杂胞(61.79)
	5-氮杂胞(25.84)
	5-氮胞苷(12.35)

cytidine deaminase	胞苷脱氨酶(100)
cytidine monophosphate	胞苷一磷酸(100)
cytidine nucleotides	胞苷核苷酸(100)
cytidine triphosphate	胞苷三磷酸(100)
deoxy cytidine	脱氧胞苷(100)
deoxycytidine	脱氧胞苷(100)

Table 3 Fuzzy Search Results Compared (source: PatentLex)

In Table 4 below, authentic examples of usage from different domains of patents according to PCT classifications are provided from PatentLex, but not from HOWNET, because they are not available.

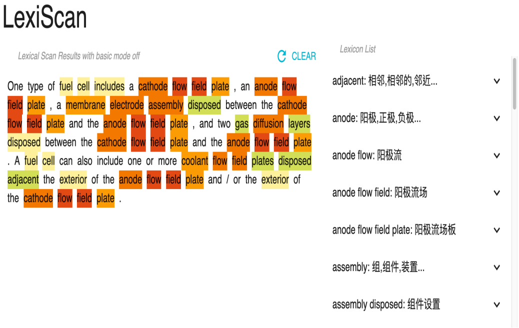
IPC	English	Chinese
C07	GDMEM contains DMEM (Gibco) and 4.5g/l glucose, 15 mg/1 phenol red, 1 mM sodium pyruvate, 1.75 g/1 sodium bicarbonate, 500 µM asparagine, 30 µM adenosine, 30 µM guanosine, 30 µM **cytidine**, 30 µM uridine, 10 µM thymidine, and non-essential amino acids (GIBCO).	GDMEM 培养基含 DMEM (Gibco)， 4.5g/L 葡萄糖， 15mg/L 酚红， 1mM 丙酮酸钠， 1.75g/L 碳酸氢钠， 500 µm 天门冬酰胺， 30 µm 腺苷， 30 µm 鸟苷， 30 µm **胞苷**， 30 µm 尿苷， 10 µ m 胸腺嘧啶核苷和非必需氨基酸（GIBCO）。
A61	The cells were pre-incubated for 27.5 hours in **5-azacytidine** before addition of SAHA.	将细胞在 **5-氮杂胞苷**中预温育 27.5 小时， 然后加入 SAHA 。
C07	The short plasma half-life is due to rapid inactivation of decitabine by deamination by liver **Cytidine deaminase**.	短的血浆半衰期是由于肝脏**胞苷脱氨酶**通过脱氨基作用对地西他滨的快速灭活而导致 。

Table 4 Some authentic example sentences for the alternate Chinese renditions (source: PatentLex)

It can be seen from the case of *cytidine* that there are considerable quantitative and qualitative differences between HOWNET and PatentLex. They show how the cultivation of a specialist corpus could expedite the user’s search and so enhance his productivity by optimizing his search.

2.8 Lexical Scan

Apart from being able to enjoy the useful feature of cross-lingual search engine which is helpful not just for translation, it is usual for a translator to face two challenges when tackling a new piece of text: (a) new technical terms not in his vocabulary, and (b) making a proper selection where there are multiple renditions. It would be very helpful if he or she is given some relative weighting (i.e., relative frequency of usage) of the alternate renditions, and if he or she could review actual examples from the technical texts where necessary. For this reason, a text scanning feature has been added by drawing on all bilingual multi-word expressions extracted in the previous effort. It includes (a) *renditions lookup*, (b) *distribution profiling*, (c) *authentic text lookup*, (d) *thesaurus navigation* in an integrated interface. An example of semantic net navigation of LexiScan involving (a) and (b) is given below:



LexiScan

One type of fuel cell includes a cathode flow field plate , an anode flow field plate , a membrane electrode assembly disposed between the cathode flow field plate and the anode flow field plate , and two gas diffusion layers disposed between the cathode flow field plate and the anode flow field plate . A fuel cell can also include one or more coolant flow field plates disposed adjacent the exterior of the anode flow field plate and / or the exterior of the cathode flow field plate .

- adjacent: 相邻,相邻的,邻近...
- anode: 阳极,正极,负极...
- anode flow: 阳极流
- anode flow field: 阳极流场
- anode flow field plate: 阳极流场板
- assembly: 组,组件,装置...
- assembly disposed: 组件设置

The LexiScan produces lexicon list which includes a full range of related terms in the sidebar:

adjacent: 相邻, 相邻的, 邻近
anode: 阳极, 正极
anode flow: 阳极流
anode flow filed: 阳极流场
anode flow filed plate: 阳极流场板
assembly: 组, 组件, 装置
assembly disposed: 组件装置

Given a piece of technical text, LexiScan will highlight all recognized technical terms from the PatentLex database, together with their renditions on the sidebar as shown (*renditions lookup*). Since a single word may be part of multiple multi-word expressions, a darker shade indicates that the single word contributes to other multi-word expressions. In the above LexiScan example, if we click on the word "*flow*" in the phrase "*cathode flow field plate*" on the first line, we can view the following list of words containing the word "flow" in English but not necessarily with the same renditions in Chinese.

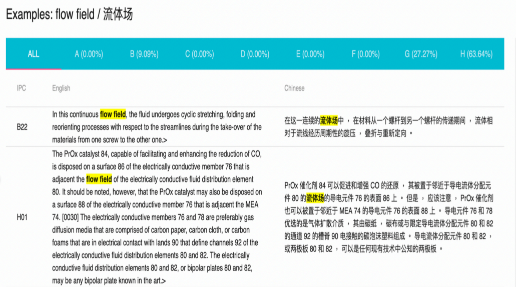
a. cathode flow: 阴极流
b. cathode flow field: 阴极流场
c. cathode flow field plate: 阴极流场板
d. flow field: 流场, 气流场, 流动区
e. flow field plate: 流场板

From the above, we can thus see a number of Chinese terms providing different renditions of "flow": namely "*cathode flow*", "*cathode flow field*", "*cathode flow field plate*", "*flow field*", and "*flow field plate*". They are all possible multi-word expressions covering "flow" from the lexicon. We could further expand some of the expressions on the list to examine the distribution of the possible renditions (*distribution profiling*).

Lexicon list (filtered) for flow field

a. 流场（94.94%）
b. 气流场（2.04%）
c. 流动区（1.69%）
d. 流动场（0.95%）
e. 流体场（0.35%）
f. 流动域（0.03%）

Furthermore, useful authentic bilingual example sentences drawn from patents may be obtained by clicking on "flow field" and "流体场" (*authentic text lookup*).



Examples: flow field / 流体场

IPC	English	Chinese
B22	In this continuous flow field, the fluid undergoes cyclic stretching, folding and reorienting processes with respect to the streamlines during the take-over of the materials from one screw to the other one.>	在这一连续的流体场中，在材料从一个螺杆到另一个螺杆的传递期间，流体相对于流线经历周期性的旋压，叠折与重新定向。
H01	The PrOx catalyst 84, capable of facilitating and enhancing the reduction of CO, is disposed on a surface 86 of the electrically conductive member 76 that is adjacent the flow field of the electrically conductive fluid distribution element 80. It should be noted, however, that the PrOx catalyst may also be disposed on a surface 88 of the electrically conductive member 76 that is adjacent the MEA 74. [0030] The electrically conductive members 76 and 78 are preferably gas diffusion media that are comprised of carbon paper, carbon cloth, or carbon foams that are in electrical contact with lands 90 that define channels 92 of the electrically conductive fluid distribution elements 80 and 82. The electrically conductive fluid distribution elements 80 and 82, or bipolar plates 80 and 82, may be any bipolar plate known in the art.>	PrOx 催化剂 84 可以促进和增强 CO 的还原，其被置于邻近于导电流体分配元件 80 的流体场的导电元件 76 的表面 86 上。但是，应该注意，PrOx 催化剂也可以被置于邻近于 MEA 74 的导电元件 76 的表面 88 上。导电元件 76 和 78 优选的是气体扩散介质，其由碳纸，碳布或与限定导电流体分配元件 80 和 82 的通道 92 的槽脊 90 电接触的碳泡沫塑料组成。导电流体分配元件 80 和 82，或两极板 80 和 82，可以是任何现有技术中公知的两极板。

As in cross-lingual search, the percentage distribution across patent domains can be seen from the headings. Bilingual examples belonging to a specific patent class can similarly be narrowed down by clicking on the heading. It is worth noting that the LexiScan feature has wider application in cross-lingual translation, especially for textual analytics.

These are recognized as useful provisions for professional translator as well as for students and teachers of translation.

2.9 Other Features Analysis

Additionally, there are wider applications, especially for cross-language patent search and analysis via the production of knowledge graphs which could help the user to navigate through relevant research networks (Tsou et al., 2019). Moreover, an intermediate stage during the 10 years long curation process has included the harvesting of a very sizable corpus of bilingually aligned sentence pairs useful for training and evaluating Chinese-English machine translation engines (Lu et al., 2011b; Goto et al., 2012, 2013; Tsou et al., 2017a, 2018).

2.10 Concluding Remarks

In the Age of Big Data, there is easy availability of data, but this ease in availability should not be mistaken for ease in securing quality. Considerable care and efforts must go into the cultivation and evaluation of the data for its full value to be realized. This paper has discussed how a database of 1 million entries of highly valued bilingual multi-word expressions in the technical fields has been profitably mined from a combined database involving 10 years of English and Chinese patents, comprising 5,353 million English words and 12,000 million Chinese characters. The processes involved and the utilization of the resultant database and platform, PATENTLEX, have been outlined. It is hoped that the case reported here has demonstrated the considerable value of well curated corpus not just for mining lexical gems, but for other data mining as well.

References

Gale, William A., and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In Proceedings of ACL. pp.79-85.

Dyer, Chris, Chahuneau, Victor and Noah A. Smith.: A simple, fast, and effective reparameterization of IBM Model 2. Proceedings of NAACL-HLT, pp. 644–648. (2013).

Adafre, Sisay Fissaha and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In Proceedings of EACL, pp. 62-69.

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In Proceedings of ACL. pp.169-176.

Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. In Proceedings of ACL. pp. 9-16.

Goto, Isao, Bin Lu, Ka-Po Chow, Sumita Eiichiro, and Benjamin K. Tsou. (2012). “Overview of the Patent Translation Task at the NTCIR-9 Workshop”. In Proceedings of the NTCIR-9 Workshop, pp. 559-578. Tokyo.

Goto, Isao, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou.: Overview of the patent machine translation task at the NTCIR-10 workshop. Proceedings of NTCIR-10 Workshop Meeting. (2013).

Och, Franz J., and Hermann Ney. 2004. The Alignment Template Approach to Machine Translation. Computational Linguistics, 30(4), 417-449.

Jiang Long, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. In Proceedings of ACL-IJCNLP. pp. 870-878.

Tian Liang, Fai Wong, and Sam Chao.: Phrase Oriented Word Alignment Method. In Wang, Hai Feng (Ed.), Proceedings of the 7th China Workshop on Machine Translation pp. 237‑250. Xiamen, China (2011).

Tian Liang, Derek F. Wong, Lidia S. Chao, and Francisco Oliveira.: A Relationship: Word Alignment, Phrase Table, and Translation Quality. The Scientific World Journal1–13 (2014).

Bin Lu, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, and Olivia Y. Kwong. (2009). The Construction of an English-Chinese Patent Parallel Corpus. MT Summit XII 3rd Workshop on Patent Translation.

Bin Lu, Benjamin K. Tsou, Tao Jiang, Oi Yee Kwong and Jingbo Zhu. 2010a. Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. In Proceedings of the 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2010). Beijing, China.

Bin Lu, Ka-po Chow and Benjamin K. Tsou. 2011a. The Cultivation of a Trilingual Chinese-English-Japanese Parallel Corpus from Comparable Patents. In *Proceedings of Machine Translation Summit XIII (MT Summit-XIII)*. Xiamen.

Bin Lu, Benjamin K. Tsou, Tao Jiang, Jingbo Zhu, and Kwong, Olivia. 2011b. “Mining parallel knowledge from comparable patents”. In Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances. pp. 247-271. IGI Global.

Bin Lu, Ka-po Chow and Benjamin K. Tsou (2015). “Comparable Multilingual Patents as Large-scale Parallel Corpora”. In: Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, Pascale Fung (Eds.), Building and Using Comparable Corpora. Springer-Verlag, pp 167-187.

Bin Lu, Benjamin K. Tsou and Ka-po Chow. (2016). Cultivating Large-scale Parallel Corpora from Comparable Patents: From Bilingual to Trilingual, and Beyond. In Tsou, Benjamin, and Kwong, Olivia., (eds.), Linguistic Corpus and Corpus Linguistics in the Chinese Context (Journal of Chinese Linguistics Monograph Series No.25). Hong Kong: The Chinese University Press, pp. 447-471.

Xiao-yi Ma.: Champollion: A Robust Parallel Text Sentence Aligner. In Proceedings of the 5th

International Conference on Language Resources and Evaluation (LREC). Genova, Italy (2006).

Utiyama, Masao, and Isahara Hitoshi. 2007. A Japanese-English patent parallel corpus. In Proceeding of MT Summit XI. pp. 475–482.

Utiyama, Masao and Isahara Hitoshi.: Reliable measures for aligning Japanese-English news articles and sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 72– 79. Sapporo, Japan (2003).

Simard, Michel, and Plamondon Pierre. 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. Machine Translation, 13(1), 59-80.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In Proceedings of MT Summit X.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of ACL Demo Session. pp. 177-180.

Koehn, Philipp.: Statistical machine translation. Cambridge University Press, the United Kingdom (2010).

Resnik, Philip and Smith Noah A. (2003) The Web as a Parallel Corpus, Computational Linguistics 2003 29:3, pp. 349-380

Haldar, Rishin and Mukhopadhyay Debajyoti. “Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach.” ArXiv abs/1101.1232 (2011).

Jason R. Smith & Quirk, Chris & Toutanova, Kristina. (2010). Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. 403-411.

Sharoff S., Rapp R., Zweigenbaum P. (2013) Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In: Sharoff S., Rapp R., Zweigenbaum P., Fung P. (eds) Building and Using Comparable Corpora. Springer, Berlin, Heidelberg.

Smith, Jason R., Chris Quirk and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In Proceedings of NAACL-HLT. pp. 403-411.

Benjamin K. Tsou, Bi-wei Pan, and Ka-po Chow. Some Challenges and Advances in Natural Language Processing of Chinese Patents ─ From Machine Translation to Cognitive Filtering. [Keynote paper], WIPO East Meets West Seminar. Vienna (2017a).

Benjamin K. Tsou, Derek Wong, and Ka-po Chow. Successful Generation of Bilingual Chinese-English Multi-word Expressions from Large Scale Parallel Corpora: An Experimental Approach, paper presented at EUROPHRAS. London (November 2017b).

Benjamin K. Tsou, Min-yu Zhao, Bi-wei Pan, and Ka-po Chow.: The Age of Big Data and AI: Challenges and Opportunities for Technical Translation 4.0 and Relevant Training. Translators Association of China (TAC) Conference. Beijing (2018).

Benjamin K. Tsou, Ka-po Chow, Jun-ru Nie and Yuan Yuan: Toward a Proactive MWE Terminological Platform for Cross-Lingual Mediators in the Age of Big Data. Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT) (2019).

Germann, U. Aligned Hansards of the 36th Parliament of Canada (2001), http://www.isi.edu/natural-language/download/hansard/

Dekai Wu, and Pascale Fung 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In Proceedings of IJCNLP2005.

Xiaodong Zeng, Lidia S. Chao, Derek F. Wong, Isabel Trancoso, and Liang Tian.: To- ward Better Chinese Word Segmentation for SMT via Bilingual Constraints. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1), pp. 1360–1369. Baltimore, Maryland (2014).

Bing Zhao, and Stephen Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In Proceedings of Second IEEE International Conference on Data Mining (ICDM-02).

Functional Text Representations for Building Cross-Linguistic Comparable Corpora in English and Russian

Maria Kunilovskaya
University of Wolverhampton
maria.kunilovskaya@wlv.ac.uk

Shiva Taslimipoor
University of Wolverhampton
shiva.taslimi@wlv.ac.uk

Tatyana Ilyushchenya
University of Tyumen
t.a.ilyushhenya@utmn.ru

Abstract

In this contribution we report the results on cross-linguistic building of functionally comparable corpora. Functional similarity of corpus resources is an important prerequisite for translationese studies, which traditionally reveal translations as texts deviating from the conventions of the intended genre in the target language. Therefore, measuring translationese is directly contingent on the corpus of non-translated target language selected to represent the expected norm for a given genre. Functional similarity of the corpora is also key for contrastive analysis. We propose a solution based on representing texts with functional vectors and comparing texts on these representations. The vectors are produced by a recurrent neural network model trained on the hand-annotated data in English and Russian from the Functional Text Dimensions project. Our results are verified by an independent annotation experiment, and the tests run on an evaluation corpus. The latter experiments are set to investigate whether the vectors capture traditionally recognised genres and the expected cross-linguistic degree of text similarity. We apply this approach to describe the functional similarity of the 1.5 million token English and Russian subsets of the respective hundred-million word Aranea corpora, the comparable web-corpora project.

1 Introduction

Corpus-based translation studies and contrastive analysis typically require intra- and inter-linguistically comparable corpora. The comparability of the resources is usually ensured by collecting texts from similar sources (e.g. the same institutions, websites, or corpora), and by using the same chronological and sociolinguistic sampling frame. Alternatively, researchers can rely on the pre-existing register/genre annotation. Sometimes, the description of the resources comparability is limited to a phrase such as ‘the BNC sample was chosen so as to mirror the makeup of the TEC’ or ‘reference corpus made comparable to the parallel data in terms of register’. The assumed comparability of monolingual and cross linguistic resources is typically a point of criticism. For example, in his overview of research on explicitation, Becher (2011) questions the comparability of materials used in numerous cases. The importance of building an adequate reference corpus is also reflected in the fact that some corpora (like CroCo) that are designed for translationese or contrastive research, include the untranslated reference texts as their integral part (Hansen-Schirra et al., 2012). It is a well known fact that different registers/genres trigger different type of translationese: Lapshinova-Koltunski (2017) shows that register is one of the major factors explaining variation in translation along with translation method and expertise. Neumann (2013) revealed the specificity of German-English translations observed in some registers but not others.

The above demonstrates that the concept of corpus comparability in translation studies or contrastive analysis is not based on the domain or ‘aboutness’ of the texts, but has to do more with the ‘context of situation’. It is the interplay of various parameters of the communication event that are important for defining genres. Therefore, despite most research in corpus comparability defining comparable corpora as texts in the same topic domain — e.g. they are harvested on a set of seed terms (Kilgarriff et al., 2011); comparability is calculated based on the lexical features, such as vocabulary overlap or bag-of-words representations (Li et al., 2018). This research interprets comparability as a functional or genre-related property, similarly to how bilingual comparable corpus is described in Kutuzov et al. (2016), or how it is traditionally defined in corpus-based translation studies (Zanettin, 2012).

This paper aims to test whether abstract and language-independent functional properties of texts can be used as a text-external approach to cross-lingual text categorisation. Namely, we explore the usability of the Functional Text Dimensions, a set of text functions hand-annotated for English and Russian web texts (Sharoff, 2018), as a training data to produce vectorised representations of texts functionality. Text functions, which reflect the speaker's communicative goal, are among of the major descriptors of a communicative event and are invariably present in the genre definition. Besides, it is one translationally relevant aspect of texts that can be used to build cross-lingual comparable resources for translationese studies.

In addition to the intrinsic evaluation of the models' performance, we provide results of the external evaluation in two aspects. First, we evaluate the effectiveness of the functional vectors for genre classification against alternative text representations. For these purposes we use a selection of 'known' genres extracted from the national corpora in the two languages. Second, the cross-linguistic comparability of the models' output is tested by measuring the average functional similarity of text pairs coming from subcorpora with varying degrees of similarity.

The rest of the paper is structured as follows. Section 2 outlines the research on genre identification and text functionality that we draw upon. In Section 3 we describe our training data, the settings of the modelling experiment, including the architecture of the recurrent neural network model, as well as experimental results. Then, we predict functional vectors for our evaluation corpora and estimate these vectors against expected standard in Section 4. Section 5 has a brief description of the application of the functional vectors to the description of the English and Russian samples of the Aranea web-corpora. The final section (Section 6) summarises the results.

2 Related Research: Register Studies

Apart from the domain-based text categorisation typical for NLP tasks, there are two major approaches to describe text variation in register/genre studies. The text-*internal* approach to text categorisation is based on calculating frequencies of lexicogrammatic features ('register' features, such as conjunctions, passives, modals, pronouns, tenses), that allegedly reflect linguistically relevant parameters of the communicative situations. One of the best known implementations of this approach is Biber's work (Biber, 1988).

The text-*external* approach draws on the audience's perception of the author's communicative aims and known circumstances of the text production (the author's social role, mode of speech, degree of the participants' interaction), and uses genres as a loose set of culture-specific categories to explain text variation.

There is no arguing that these views are complementary. Calculating frequencies of tokens (lexis-based catagorisation typical for domain-oriented approach), can be as effective in genre classification as the more elaborate register features. Xiao and McEnery (2005) show that keywords analysis can be as effective in detecting both similar (everyday conversation vs official speech) and distant genres (spoken genres vs. academic prose) as Biber's features.

There have been also numerous attempts to establish a link between genres and their linguistic features, while ignoring domain differences inside genre categories (including Lee and Myaeng (2002) and Braslavski (2010)). However, the researchers have to use a pre-existing genre typology, which pigeonholes texts in accordance with the accepted convention in the given language community, and does not allow for a more flexible and realistic reflection of the evolving text-type variety or for reliable cross-linguistic comparisons. Moreover, simple solutions, which work for the major text categories, fail in the presence of more subtle distinctions. For example, we have found that the impressive and reproducible results from Lijffijt and Nevalainen (2017), where they achieved F1 = 90% in the classification of the four 'tried and tested' top-level categories from BNC using pairs of the simple register features like frequencies of nouns and pronouns, gets reduced to only F1 = 71% on a less balanced six categories subcorpus described in Kunilovskaya and Sharoff

(2019).

One approach to avoid the atomic genre labels and to work around the culture-specific nature of genre categorisation is to represent texts as vectors in a multi-functional space, where a text can get relative scores on several dimensions. An attempt to define such a space using the speakers' perception of the candidate text proximity to the recognizable functional prototypes was made in the Functional Text Dimensions (FTD) annotation project presented in detail in Sharoff (2018). This framework is particularly appealing for our purposes because it is based on translationally relevant functional properties of texts and offers a theoretically reasonable tertium comparationis required for cross-linguistic corpus building.

3 Modelling Functional Text Representations

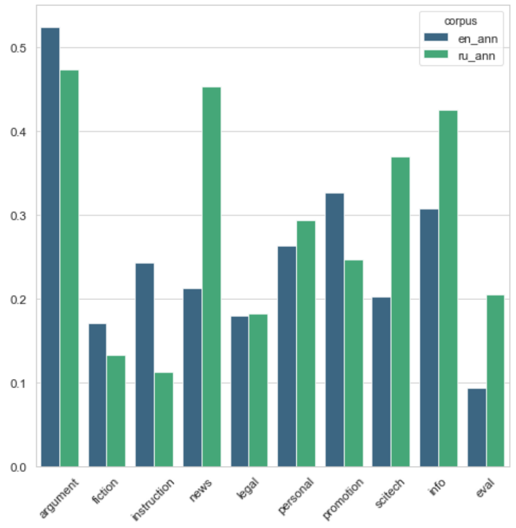
In this section we describe the annotated data in English and Russian from the FTD project and illustrate the neural network approach that we used for modelling text functionality.

3.1 Data

The annotated data consists of 1,624 chunks of texts that count about 2 million tokens for English; the Russian part of the project includes 1,930 texts (2.4 million tokens). For both languages the texts come from two sources: 5gthe Pentaglossal corpus (Forsyth and Sharoff, 2014) and ukWac (Baroni et al., 2009). We used the annotations for the ten most prominent FTD described in Sharoff (2018). Sharoff (2018) also has a detailed description of the original annotation experiment and reports the inter-annotator agreement at Krippendorff's alpha >.76. The annotators were asked to score each text on the 4-point Likert scale (0, 0.5, 1, 2) depending on how much the text resembles the suggested functional prototype for each dimension. While referring the reader to the original paper for more details, a few examples of the labels and prototype texts used in the annotation project are: A1 (argument) blogs, editorials, opinions; A7 (instruction) tutorials or FAQ; A8 (hardnews) report of events, inc. future events.

The original dataset was augmented by splitting longer texts into additional instances. The text length used for training was set to 1000 words. This re-sampling helped the distribution of the FTD labels to be more even than in the original dataset and normalized the text length. Figure 1 depicts the distribution of texts assigned to different functions by annotators for both English and Russian. The corpora of the two languages follow the same building frame: the data come from the same multilingual resources, which makes it possible to assume their comparability.

Figure 1. How often each function receives a positive score in the annotated data (proportion to all texts)



3.2 Model

Our model is a bidirectional LSTM with an attention layer on top. The input to the model are the embeddings of the words in the text, pretrained on the Common Crawl data from the fastText project (Grave et al., 2018). The output is a 10-dimensional functional vector for each document, namely, the functional representation of the document. In this experiment the model was set up to recognise the functions manifested in the text, rather than learn the scores assigned by the annotators. To this end, the annotations for each FTD were binarised (0.5 is set to 0, and 2 is set to 1). Consequently, we had a multi-hot vector for each document as our target.

As a simple baseline, we used a classifier which attempted to learn the binarised values for each FTD separately. In addition we set up a multi-task learning scenario in which the model learns all 10 binary values simultaneously. In this case, our learning model back-propagated based on the accumulation of the loss functions for all 10 labels. In another experiment we enriched the embedding features with the Biber's register features of the documents. For extraction of these features we relied on MAT for English (Nini, 2015), and the framework provided in Katinskaya and Sharoff (2015) for Russian.

3.3 Results

For all our models, we used CuDNN LSTM and trained the model for 20 epochs. We used Adam optimizer and 0.2 dropout after embedding and 0.5 dropout after the LSTM layer. The loss function was binary cross entropy since we predict binarised valued of FTD columns.

In Table 1 we report the performance measures on the 10-fold cross validation for the main models (biLSTMa), and the models which use the Biber's features together with embeddings (biLSTMa-bib) – both in multi-task settings, compared to the baseline (biLSTMa-10b) for both languages.

We evaluate the performance of the models in predicting FTD values in three ways: first, we established how well they predict individual functions on average. Second, given the sharp imbalances between the positive and negative classes for each function, we report the F1 measure for the minority class. Finally, the last two columns in Table 1 have the F1 and accuracy statistics for individual samples. Since our target is a multi-hot 10-component vector, accuracy (which counts an observation as correctly predicted only if all the 10 classes are correctly classified) is very strict. Instead, we opt for the negative hamming loss, i.e. the ratio of all correctly predicted labels for an instance to all labels. To deal with severe class imbalances, we use stratified (multi-label) split with cross-validation and at the evaluation stage by we choose macro-averaging which penalises model errors regardless of class distributions. The results in Table 1 show that our multi-task architectures (rows indicated with multi) outperform the baseline in which the 10 values for each text are learned independent of each other. The better performance of adding register features to our model can be seen only in the case of Russian.

These results show the effectiveness of our models in estimating the probability that a text fulfills the corresponding functions. and leads us to further use these vectors as functional representations for text. In the next two sections we first demonstrate how the predictable functional representations correlate with the text's general functional and genre properties which are external to the initial annotation experiment. We also demonstrate the application of the functional vectors for corpus comparison in Section 5.

4 External Evaluation

4.1 Functional vectors for BNC/RNC genre categories

To determine whether our functional representations are useful to distinguish the text categories outside the original annotated corpora, we designed a 'known' genre composition corpus. To compile this genre evaluation corpus, we used the metadata in the British National Corpus (BNC) and the Russian National Corpus (RNC), described in Lee (2001) and Savchuk (2006) respectively. We focused on the genre categories which approximate some of the prototype texts described in the annotation guidelines and are annotated in both national corpora. We extracted the written texts that were longer that 400 words by the tags (in the order shown in Table 2).

For English we extracted all texts tagged as follows: *ac:nat science, fict prose, nonAc: nat science, newsp brdsht nat: report, newsp other: report, biography, advert*. For Russian the texts for each category were selected by the tags combinations: academic (sphere=science and education, type=article, topic=science and technology, audience != big, level=professional/high), fiction (sphere=fiction, type=short story, story, novel), reportage (sphere=publicist, type=info message), personal (sphere=publicist, type=memoirs/biography), promotion (sphere=promotion). [1] Unlike the BNC, the RNC has no separate text type for non-academic texts. To remedy this incompatibility, we used chapters from 14 popular Russian scientific books in academic domains such as linguistics, biology and anthropology published between 2010 and 2017. The books were split into 1000-word chunks; a random selection of 100 of those chunks was used as part of the Russian evaluation corpus. The resulting collections were balanced in terms of the number of texts per each category: we retained 100 random documents for

[1] For Russian we additionally limited the sampling frame to include only the texts published after 2004, neutral of style and marked as intended for a large audience, with no restriction by age or education level (with the exception for academic texts).

		FTD overall			FTD minority	Samples	
		P	R	F1	F1	F1	hamming loss
baseline	biLSTMa-10b (EN)	.810	.853	.824	.655	.683	.927
	biLSTMa-10b (RU)	.799	.878	.825	.644	.709	.924
multi	biLSTMa (EN)	.824	.862	.841	.722	.483	.930
	biLSTMa (RU)	.818	.871	.841	.724	.504	.922
multi+bib	biLSTMa-bib (EN)	.814	.861	.835	.711	.472	.927
	biLSTMa-bib (RU)	.829	.875	.849	.742	.522	.926

Table 1. Results for FTD modelling experiments

each category that counted more than 100 texts. We further truncated the texts to the first 1000 words, if the selected texts were longer. Table 2 shows the basic parameters of the BNC and RNC subcorpora used for evaluation in this study. We also include the FTD which is expected to be dominant in the texts of each category.

To investigate the reliability of the predicted functional vectors for genre analysis, we classified the texts into six categories that are listed in Table 2. We report the results received in the same settings for the alternative text representations, namely, the raw Biber's features and the keywords statistics.

The Biber's features were extracted with MAT and MDRus analyzer (Nini, 2015; Katinskaya and Sharoff, 2015). The keywords features for each text were calculated using the log likelihood (LL) measure against all of the data from the respective national corpus used in this experiment. Prior to keyword extraction the data was lemmatized, and functional words were filtered out, leaving us with a vocabulary of 24k and 40k content lemmas for English and Russian respectively. To reduce the sparsity of the data, we limited the list of keywords to the top 100 with LL >6.63 (the standard 1% significance level) for each text and, further, to only those which occurred in 3% of the texts. The number of the resulting keywords was 408 for the BNC selection and 489 for the RNC.

In Table 3 we report the macro-average stratified 10-fold cross-validation results for a RandomForest classifier with the default scikit-learn parameters (n_estimators=10, criterion='gini', bootstrap=True) (Breiman, 2001). The results show that the classification using functional representations (vectors) outperforms the classification using alternative ways of representation, given our selection of genres and the classification settings. Interestingly, the combination of functional and the Biber's features yield a 2% increase in the performance of the classifier.

It can be seen that the results for English in this experiment were consistently better than those of Russian. In-depth analysis of the classifier performance per category (omitted here for brevity) showed that the algorithm struggled with different genres for different languages. In Russian *reportage* proved to be the most challenging genre (the F1 score for predictions of items in this category was 0.56), while *fiction* returned the highest results (F1 = 0.79); In English non-academic texts were comparatively difficult to solve (F1 = 0.63), while reportage and promotion were comparatively easily recognised (F1 = 0.87 and F1 = 0.83).

4.2 Cross-linguistic comparability of English and Russian functional vectors

In this section we test whether the texts in English and Russian that are expected to be functionally similar in the real world receive similar functional representations in our experiments. We have seen that our predicted vectors are able to detect the generic properties of texts reflected in the hand-crafted text category metadata of the two national corpora. However, it is not clear whether the vectors produced by the models learnt on the English and Russian data are effective in measuring their cross-lingual comparability.

To explore this aspect of the functional representations, we measured and compared similarity between the four sets of text pairings which are expected to display decreasing degrees of functional similarity: (1) aligned parallel texts of the four genres; (2) texts from the same genres in the parallel corpus that are not translations of each other; (3) random text pairs from the comparable categories of the national corpora (described in Table 2) and (4) random text pairs for texts from

		academic	fiction	nonac	pers	promo	rep	total
BNC	texts	43	100	62	100	59	88	452
	words	38k	86k	56k	88k	47k	79k	394k
RNC	texts	100	100	100	100	100	82	582
	words	92k	94k	104k	94k	75k	52k	k 511k
FTDs		A14	A4	A1	A8	A11	A12	

Table 2. The composition of the genre-balanced comparable evaluation corpus

	EN	RU
vectors	.77	.68
Biber's	.73	.64
keywords	.66	.60
vectors+Biber's	.79	.70

Table 3. Classification results for the six categories in each national corpus selection

	texts	words
fiction	170	9.6m
media	132	133k
ted	100	259k
popsci	100	826k

Table 4. Composition of the parallel component of the evaluation corpus

different genres in the national corpora (the negative similarity material). We used the Euclidean measure of similarity to compare the functional vectors between the text pairs in each set. This measure takes into account the magnitudes of the vectors components, which are meaningful in our representation.

The highest degree of the expected cross-lingual functional similarity is represented by the professional translations and their sources. Within the functional theories of translation, which underlie the current professional norm, good translations are expected to reproduce the functional hierarchy of the source. The texts in this section of our evaluation corpus are extracted from the parallel English-Russian component of the RNC[2] (fiction, mass-media texts) (Dobrovolskij et al., 2005) and from the professional translations segment of the RusLTC project[3] (including TED talks and popular scientific texts) (Kutuzov and Kunilovskaya, 2014). The descriptive statistics for the parallel component of the evaluation corpus are given in Table 4 (the word count is based on the English sources).

In Table 5 we report average pair-wise similarity for the documents of different categories in the four aforementioned sets. As can be seen in the table, as the level of comparability decreases from the top set to the bottom set, the calculated similarity also decreases.

The fluctuation in the similarity values for the different genres (see Table 5) indicates that the translations of popular-scientific books are the most functionally faithful (see the results for aligned translations). In the comparable part of the evaluation corpus the academic texts demonstrate the highest similarity of 0.396, while the least functionally similar genres are non-academic (popular-scientific) texts (0.127), personal writings, such as biographies and memoirs (0.139), and promotional texts (0.145).

4.3 Results from an independent annotation effort

To test the generalisation power of the models on external data, we ran an independent annotation experiment, following the guidelines described in Sharoff (2018) and summarised in Section 3.1. We had three trained linguists assign 10 functional scores to each of 70 English texts selected randomly from two parallel collections: CroCo (Hansen-Schirra et al., 2006) and the RusLTC corpus mentioned in Section 4.2. After discussing the results on the first 10 texts, which revealed some differences in the task interpretation, the three raters reached the overall agreement of Krippendorff's $\alpha = 0.537$. In cases of triple disagreements (48 items out of 700), the three different values assigned by annotators were averaged and rounded to the closest score.

We used the same evaluation strategy as in Section 3.3 and compared the binarised human scores

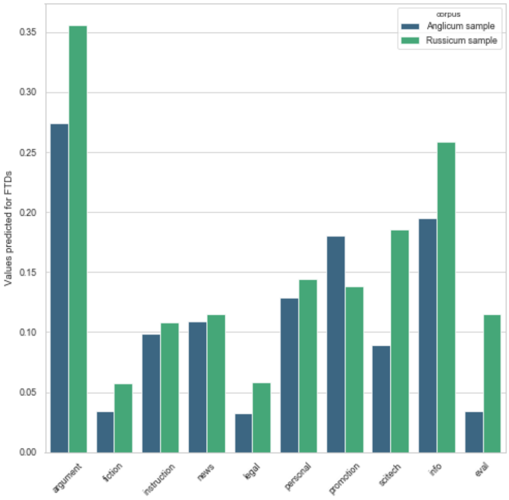
[2] http://www.ruscorpora.ru/new/corpora-structure.html

[3] https://www.rus-ltc.org/static/html/about.html

set	category	similarity	mean
aligned	fiction	.432	.470
	media	.476	
	ted	.456	
	pop-sci	.514	
unrelated	fiction	.315	.305
	media	.263	
	ted	.323	
	pop-sci	.317	
same genre	academic	.396	.214
	fiction	.259	
	non-academic	.127	
	personal	.139	
	promotion	.145	
	reportage	.216	
unrelated	academic::fict	-.190	.004
	non-ac::promo	.116	
	pers::report	.085	

Table 5. Average functional similarity measures for the four components of the evaluation corpus

Figure 2. The overview of the texts functionality based on the average values for FTDs: Anglicum Minus vs Russicum Minus



and the English model predictions. The average result over the 10 FTDs reached macro F1 score of 0.732. The lower performance of the model on this data can be explained by a different distribution of the text types in the annotated 70 texts as compared to the training set.

5 Application to Aranea

In this section we demonstrate the application of the vectors to analyse the genre composition of comparable web-corpora from the Aranea project. We randomly selected 1% of texts from the 120-million token Araneum Minus Anglicum and Araneum Minus Russicum (Benko, 2014)[4] and represented them with functional vectors. The samples include around 4.5k texts that are over 450 tokens long, and count 1 and 1.5 million tokens, respectively. Since there is no prior information on the internal generic structure of the corpora we can not measure their overall similarity directly (as we did with the national corpora). The purpose of this exercise is only to provide a comparative description of the corpora genre composition.

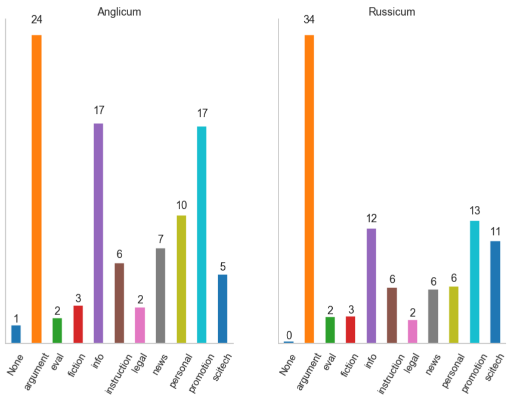
We tried two ways of capturing the contents of the corpora: 1) the average value for each function and 2) the ratio of the texts with a given function as the predicted dominant function. The dominant function is defined as the FTD with the highest probability value returned by the model for a given text. In either case the general picture comes at the price of losing the functional hierarchy and the possible hybrid nature of texts. From this perspective the Russian texts in the analysed slices of the Aranea web-corpora have higher scores for evaluative, informative and argumentative functions (see Figure 2).

In the second approach, for each text we used only the highest functional value (i.e. their dominant function) and characterised the corpora by the ratios of texts with these dominant functions. Figure 3 shows that the Russian corpus (compared to the English one) has fewer texts with the informational and promotional as dominant functions, but it has more texts that come across as primarily scientific.

The ratios seem to be more directly comparable than the averaged probabilities, but they neglect the polyfunctional nature of many texts in Aranea: 40% of the English texts and 70% of the Russian texts have the second strong prediction (we set the threshold for the ratio between the highest value in a functional vector and the second high value at 0.7). These numbers reflect the proportion of hybrid texts in the training corpus: The human subjects assigned high scores to two (or more) functions in 40% of texts in the English part of the experiment and in 53% of texts in Russian.

[4] http://unesco.uniba.sk/aranea/index.html

Figure 3. Ratio of texts per predicted dominant function for Anglicum Minus and Russicum Minus



6 Conclusions

This paper reports the experimental results on learning functional text representations for English and Russian and describes extensive tests on their cross-linguistic comparability. We used the hand-annotated data released within the Functional Text Dimensions project to train a multi-label binary classifier based on recurrent neural networks. The average performance of the classifier is estimated at F1 >0.84 for both languages.

We evaluated the quality of the functional vectors by using them to represent texts from the six comparable text categories of the British and Russian national corpora and running a simple RandomForest classifier on the resulting data. The six-class classification returned the F1-score of 0.77 and 0.68 for English and Russian respectively. This outperformed the classification results in the same settings with the alternative representations (the Biber's and the keywords features). We saw a steady increase in the quality of the genre classification when we combined our functional vectors with Biber's features.

To evaluate the cross-linguistic comparability of the models output, we measured the Euclidean similarity between text pairs with the expected various degrees of similarity. The functional vectors learnt independently by the English and Russian models for the translationally related text pairs returned the highest similarity score of over 0.45. It is a relative score which can be interpreted in the context of the scores for the text pairs that were expected to be less similar. For example, the most dissimilar text pairs – English and Russian texts from categories with different genre labels returned 0.04. These experiments show that the functional vectors are an adequate representation of the texts functionality, a major criteria for genre identification, and can be used for measuring similarity between texts in the two languages as well as for building bilingual comparable corpora.

Acknowledgments

This work has been partly supported by the Russian Foundation for Basic Research within Project No. 17-06-00107/19.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3):209–226.

Viktor Becher. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Ph.D. thesis. http://ediss.sub.uni-hamburg.de/volltexte/2011/5321/pdf/Dissertation.pdf.

Vladimír Benko. 2014. Aranea: Yet another family of (comparable) web corpora. In *International Conference on Text, Speech, and Dialogue*. Springer, pages 247–256.

Douglas Biber. 1988. *Variations Across Speech and Writing*. Cambridge University Press.

Pavel Braslavski. 2010. Marrying relevance and genre rankings: an exploratory study. In *Genres on the Web*, Springer, pages 191–208.

Leo Breiman. 2001. Random forests. *Machine learning* 45(1):5–32.

Dmitrii Dobrovolskij, Aleksei Kretov, and Sergei Sharov. 2005. Parallel corpus: the architecture and usage potentiona [Korpus parallel'nyh tekstov: arhitektura i vozmozhnosti ispol'zovanija]. In *Russian National Corpus: 20032005 [Nacional'nyj korpus russkogo jazyka: 20032005]*, Indrik, pages 45–49. http://ruscorpora.ru/sbornik2005/17dobrovolsky.pdf.

Richard Forsyth and Serge Sharoff. 2014. Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing* 29:6–22.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proc 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing at EACL*. Association for Computational Linguistics, Trento, pages 35–42.

Silvia Hansen-Schirra, Erich Steiner, Sandra Hansen, Marlene Kast, Kerstin Kunz, Karin Maksymski, and Mihaela Vela. 2012. *Cross-Linguistic Corpora for the Study of Translations*.

Anisya Katinskaya and Serge Sharoff. 2015. Applying Multi-Dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres. *The 5th Workshop on Balto-Slavic Natural Language Processing* (September):65–74. http://www.aclweb.org/anthology/W15-5311.

Adam Kilgarriff, P V S Avinesh, and Jan Pomikálek. 2011. Comparable Corpora BootCaT. *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011, 10-12 November 2011, Bled, Slovenia* .

Maria Kunilovskaya and Serge Sharoff. 2019. Towards functionally similar corpus resources for translation. In *Proceedings of RANLP 2019*. in print.

Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko, and Lyubov Ivanova. 2016. Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints. In *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*. pages 3–10. http://arxiv.org/abs/1604.05372.

Andrey Kutuzov and Maria Kunilovskaya. 2014. Russian Learner Translator Corpus: Design, Research Potential and Applications. In *Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014, Proceedings*. Springer, volume 8655, page 315.

Ekaterina Lapshinova-Koltunski. 2017. Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. *Empirical Translation Studies. New Theoretical and Methodological Traditions* pages 207–234. https://doi.org/10.1515/9783110459586-008.

David Lee. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3):37–72.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 145–150.

Bo Li, Eric Gaussier, and Dan Yang. 2018. Measuring bilingual corpus comparability. *Natural Language Engineering* 24(4):523–549. https://doi.org/10.1017/S1351324917000481.

Jefrey Lijffijt and Terttu Nevalainen. 2017. A simple model for recognizing core genres in the bnc. In *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*, University of Helsinki, VARIENG eSeries, volume 19.

Stella Neumann. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Mouton de Gruyter, Berlin, Boston.

Andrea Nini. 2015. Multidimensional Analysis Tagger (v. 1.3).

Svetlana Savchuk. 2006. The corpus of texts from the first half of the 20th century: status quo and perspectives [korpus tekstov pervoj poloviny xx veka: tekushhee sostojanie i perspektivy]. *Russian National Corpus[Nacional'nyj korpus russkogo jazyka]* 2008:27–45.

Serge Sharoff. 2018. Functional Text Dimensions for annotation of Web corpora. *Corpora* 13(1):65–95.

Zhonghua Xiao and Anthony McEnery. 2005. Two Approaches to Genre Analysis: Three Genres in Modern American English. *Journal of English Linguistics* 33:62–82. https://doi.org/10.1177/0075424204273957.

Federico Zanettin. 2012. Bilingual Comparable Corpora and the Training of Translators. *Meta: Journal des traducteurs* 43(4):616.