Beyond Translation Memories: finding similar documents in comparable corpora

Serge Sharoff
Centre for Translation Studies
University of Leeds
s.sharoff@leeds.ac.uk

Abstract

This paper presents our most recent research in the context of TTC, an EU funded research project, on using the Web to retrieve terminologically rich texts in a specific domain, and to find similar documents in such comparable corpora. The aim of this work is to provide tools for semi-automatic construction of bilingual term lists.

1 Parallel and comparable corpora

Re-use of existing translations lies at the centre of translation practice as we know it now. Translation Memories fuel the daily activities of translators, while Machine Translation engines build their translation models from collections of parallel texts (this applies to both Statistical MT, like Google, and to traditional rule-based MT systems enriched with statistics, like Systran). However, many more texts are produced in each language on a daily basis than translated by professional translators. Even when translated texts exist, they are often not available in a given application or for a given translation task. There is a particular shortage of parallel texts in the areas undergoing recent developments. This leads to great interest in utilising comparable (=less parallel) resources for translation tasks.

Informally, any collection of texts in two languages can be positioned on a cline from 'fully parallel' to 'unrelated' with several options in between:

noisy parallel texts: such texts introduce minor language-specific adaptations, e.g., an example for searching *New York* in the OpenOffice manual might be replaced with 北京 ('Beijing') in its Chinese translation;

strongly comparable texts: such texts are not designed to be translations, while they are still devoted to the same narrow topic and their producers are aware of a related text in a different language, e.g., interlinked wikipedia articles or news items concerning exactly the same specific event in different languages;

weakly comparable texts: such texts are devoted to the same topic, but they are produced completely independently, e.g., English and Chinese textbooks on designing the wind turbines, or parliamentary debates on health care from the Bundestag, the House of Commons and the Russian Duma.

This paper reports our experience from the TTC project (Translation, Terminology and Comparable Corpora),[1] which is aimed at developing tools for (1) collecting comparable corpora from the Web, (2)

[1] http://ttc-project.eu/

Table 1: English-Chinese-Russian corpora used in this study

	Wiki-En		Wiki-Ru		Crawled-En		Crawled-Ru		Crawled-Zh	
	texts	words	texts	words	texts	words	texts	words	texts	words
Nuclear	158	254014	165	130797						
Renewable	51	99960	51	47927	5762	7505765	5126	7766462	3287	12431752

measuring their degree of comparability, (3) extracting term lists, (4) aligning the terms between different languages and (5) using the resulting bilingual glossaries in CAT and MT applications.

The use of comparable texts for training MT systems usually requires noisy parallel or at least strongly comparable texts, e.g. finding nearly parallel sentences in Wikipedia articles (Adafre and de Rijke, 2006) or in the BBC News in English and Romanian (Munteanu and Marcu, 2006). However, for the task of term extraction we can use a large amount of weakly comparable texts collected by crawling the web. In the paper, we will discuss the ways to collect comparable corpora (Section 2), ways to find similar texts and to filter out less similar ones (Section 3) and ways to use this procedure for term extraction (Section 4).

2 Sources of weakly comparable corpora

For collecting comparable corpora, three strategies are possible:

1. targeted crawling of specific resources, which are known to be comparable;
2. collection of responses from search engines using parallel terms;
3. focused crawling which starts from a small number of seeds.

A variant of the first strategy is the use of Wikipedia in which the articles are linked via the interwiki links. For collecting a topically-diverse comparable corpus, Wikipedia is a good resource providing about 600,000 aligned document pairs for English-German, 350,000 for English-Russian or 140,000 for a less-resourced pair like English-Ukrainian, all figures are for the Wikipedia dumps downloaded in November 2011 (Rapp et al., 2012). However, Wikipedia does not provide suitable resources for specific domains, see data in Table 1 for a comparable English-Russian corpus in the domain of nuclear and renewable energy. It was created by selecting specific categories provided by the Wikipedia users, e.g., 'Bioenergy', 'Tidal power' or 'Wind power', and extracting the interlinked articles classified by the Wikipedia editors under these categories in either of the two languages.

This corpus is already not large, but the resources become even more sparse when we attempt to extract a multilingual corpus in this way, e.g., there are just 13 pages in the renewable energy domain shared between English, Chinese and Russian. At the same time, the relatively small Wiki corpus from Table 1 provides the possibility of testing our methods for detection of comparable texts in a realistic setup without relying on parallel corpora, which do not exhibit topical and lexical variation possible with real comparable texts within the same domain.

The second strategy is a multilingual generalisation of BootCat (Baroni and Bernardini, 2004), in which good translation equivalents are used to find and retrieve webpages returned in response to queries to a search engine. The assumption is that parallel queries consisting of several terms yield similar pages, e.g.

wind farm	风力发电厂	ветроэлектростанция
geothermal power	地热能	геотермальная энергия
hydroelectricity	水力发电	гидроэнергетика
photovoltaics	太阳能光伏	фотоэлектричество
	...	

A set of 42 parallel terms of this kind resulted in a trilingual corpus, also listed in Table 1 (Bing was used as the search engine in this experiment). The corpus is substantially bigger than the selection of articles from Wikipedia, but it is also considerably noisier: its manual inspection reveals a large number of unrelated texts. A cleaner comparable corpus can be produced by measuring comparability of its constituent texts within and across the languages.

The third strategy needs a small set of seed URLs to start crawling by following further links from the seeds. The seed URLs might be known in advance (e.g., a rich resource for texts on a particular topic) or they can be found first using Strategy 2 by starting from a number of topically relevant words. After this, the crawling process considers the topical relevance of new pages and discards them if they stray away from the desired topic. The pruning procedure in turn implies the need to measure the similarity between the seed texts and new acquisitions. This strategy was also used in TTC by means of Babouk, a specially developed topical crawler for terminology mining (de Groc, 2011).

These approaches can be very successful in creating large comparable corpora, but they do not answer the question of how similar the corpora collected in this way are. This the question investigated in our recent work presented here.

3 Measuring the similarity across languages

3.1 Feature selection

Usually the pages are compared using their textual content as the feature. Some researchers compare texts using the most frequent words (Kilgarriff, 2001), some prefer using *hapax legomena* (Patry and Langlais, 2011). In other studies we experimented with flexigrams (Forsyth and Sharoff, 2011), i.e., combinations of words with the possibility of having gaps between them. The similarity of texts by their genres can be captured by using their part-of-speech signatures (Sharoff, 2010) or by a mixed feature set, which combines the most frequent words with POS tags for less frequent words, (Baroni and Bernardini, 2006; Sharoff, 2007).

However, because of the need to determine the similarity between terminologically rich texts in this study we restricted the feature vector for each text to the keywords extracted using the log-likelihood (LL) score. It is calculated by taking into account the relative ratio of the term usage in a document and in the rest of the corpus, as well as the absolute frequency of its occurrence, at the same time, to estimare its statistical significance as a keyword for this text (Rayson and Garside, 2000). Examples of keywords are given in Table 2, a longer English article (with the length of 1464 words) gives a longer keyword list in comparison to the Russian one (218 words only).

3.2 Anchor-based method

One way of comparing corpora across languages is by translating the features obtained from their documents and by measuring the difference in the frequencies of the translations of the keywords for each document. This can be done by computing the cosine similarity score between the feature vectors (Su and Babych,

Table 2: Keywords extracted from the articles for 'Darrieus wind turbine'

LL	DocF	Word	CorpusF	LL	DocF	Word	CorpusF
326.11	15	Darrieus	50	145.18	10	ротор (rotor)	1889
224.29	21	blade	20269	130.67	9	дарье (Darrieus)	1699
208.15	19	turbine	15976	65.57	3	самозапуск (self-start)	13
135.47	18	wind	84724	56.41	6	поток (flow)	14719
95.85	9	torque	8821	51.51	6	крыло (aerofoil)	22180
79.8	6	aerofoil	1559	34.36	2	подъёмная (lifting)	99
68.83	9	angle	39843	33.39	3	турбина (turbine)	3092
66.15	17	design	516317	29.94	3	вектор (vector)	5503
61.91	6	rotor	6944	27.29	4	скорость (speed)	35960
54.37	7	spin	29185	24.16	2	мгновенный (instantaneous)	1280
50.57	8	generate	69293	18.47	4	сила (force)	111062
47.59	6	rotate	23012	17.67	2	вращение (spin)	6518
46.95	9	speed	137005	17.25	2	плохой (difficult)	7239
35.88	4	propeller	9107	17.25	2	коэффициент (rate)	7253
35.74	2	self-starting	52				
34.51	6	pitch	69433				
34.08	3	airflow	2059				
33.51	10	force	405295				
32.11	6	tower	85230				
25.84	4	load	32271				
25.74	4	conventional	32674				
25.48	4	vertical	33781				
20.02	3	axis	21728				

2012) or by a more complicated measure, which also takes into account the salience of terms and the possibility of multiple translations (Li and Gaussier, 2010).

The drawback of the dictionary-based approaches is that they are reasonably successful only when they have a good dictionary to map the features. Even the coverage of texts by words in this dictionary is not a good approximation, since proper translations of keywords specific to the domain can be quite different from their use in a general-purpose dictionary, e.g., *blade, pitch* or *spin* from Table 2. After all, the purpose of our corpus collection procedure is to create a good glossary when it did not exist before.

With this in mind, we developed another approach, which is based on the idea of applying the same similarity measure we use within a language to gage how far the texts in question are from some known benchmarks of cross-lingual similarity. We refer to such texts as "anchor" texts. To formalise this intuition, we can represent a text by using the scores of its similarity to each of the anchor texts in the same language. We perform this operation monolingually for each language in turn. After that, the vectors of monolingual similarity to the anchor texts provide a basis for comparing documents across languages: if two documents in different languages L1 and L2 are monolingually similar to a document having two versions in L1 and L2 (e.g, an original text and its translation), we can assume the two documents are likely to be similar across the languages.

There are several options for choosing the anchors:

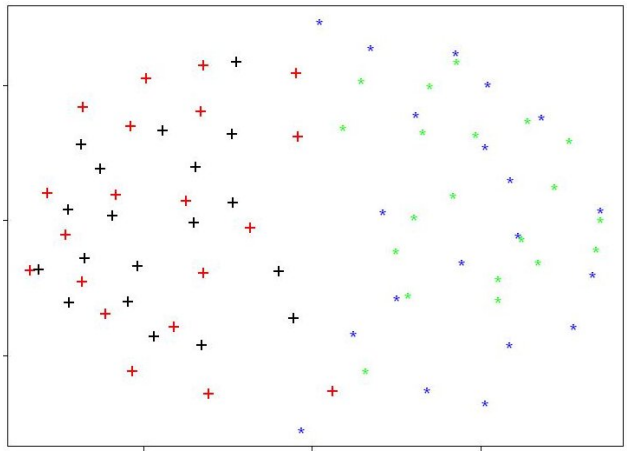
1. They can be manually selected from the crawled corpora as actually comparable texts attested by

human evaluators. This option does not require additional texts or tools apart from considerable time required to check a large number of crawled sources.

2. The anchors can be pre-selected automatically by choosing the most representative texts, e.g., the centroids of automatically produced clusters or topical models (Blei et al., 2003). A small number of centroids can be then matched manually without the need to check many potential anchor candidates. However, we have not been able to obtain reliable results using this procedure.
3. New parallel texts can be added to the crawled corpus. A small amount of parallel texts can be found for any domain, but this amount is usually not large enough for full-scale terminology extraction (often better terminological coverage is needed than available in a small number of parallel texts).
4. An array of existing parallel texts can be used as the acnhors, even when they are not exactly in the same domain as the crawled corpus. At the very least, comparability measurement against them can be used to filter out a number of texts which do not have reasonable matches in the opposite language side of the crawled corpus.

For the last option, we prepared a pentaglossal (5g) corpus, which we have used as the default for many corpus comparability studies involving Chinese, English, French, German and Russian (the five languages of the TTC project). This corpus contains 113 texts originally written in one of those languages, and later translated into all the others, such as the UN texts, TED.com transcripts, Wikinews texts, software instructions, several argumentative texts and fiction. An attempt was made to select the sources with permissive licenses, so that any wider distribution of the pentaglossal corpus is not hampered.

The similarity matrix of two corpora can be visualised using Multi-Dimensional Scaling (Sammon, 1969). The procedure attempts at mapping a set of objects from a high-dimensional space to a space with fewer dimensions (e.g., two for presentation purposes), while trying to minimise the "stress" of this transformation, i.e., the cumulative difference for the distances between all the objects in the new low-dimensional space and the same distances in the original space.



The figure illustrates this procedure by presenting the location of the Wikipedia texts from Table 1 using the 5g corpus texts as the anchors. The texts on nuclear energy are represented by pluses (black for English, red for Russian), those on renewable energy are represented by asterisks (green for English, blue for Russian). Bear in mind that the procedure did not use any dictionary apart from what it learned about the similarity of these texts to a very general corpus.

4 Prospects of using the tool for terminology alignment

The technologies for measuring the similarity of texts across languages offer particular benefits in term management. Comparable corpora are also used in Statistical Machine Translation, but their use tends to be limited to detection of parallel or nearly parallel sentences (Adafre and de Rijke, 2006; Munteanu and Marcu, 2006). On the other hand, automatic terminology extraction needs relatively large collections of texts to obtain statistically significant results. However, for alignment purposes the texts also need to be on the same topic, even at the expense of the corpus size (Morin et al., 2007). Fortunately, the process of alignment of two monolingual lists coming from two languages of a comparable corpus can ignore the exact parallelism of their syntactic constructions, and can benefit from a considerable amount of on-topic weakly comparable texts.

Below I describe our work in progress on using filtered comparable corpora to align their term lists. The alignment procedure was tested on Wikipedia with alignment information coming from the interwiki links (Rapp et al., 2012). However, it can also use a tentative mapping between the most similar documents identified using the anchor- or keyword-based approaches.

Alignment of single- and multi-word expressions in this framework is achieved without any seed bilingual dictionary (in the same spirit as the anchor-based document similarity measurement). It starts with extracting terms from the pairs of texts which have been detected as the most similar ones, using the same termhood measure as discussed above in Section 3.1. The terms in the respective lists are linked to each other with some initial weight, which can be assigned as a fixed value or it can be derived from the respective LL-score values. The subnetworks from all individual pairs are joined together for the entire set of selected document pairs. After that, activation spreads in the network using Hebbian learning, see (Rapp et al., 2012) for more details on this procedure.

In the original Wikipedia articles, only a small fraction of keywords can be identified as possible translation equivalents, see examples in Table 2. However, the procedure of spreading activation in the network is robust to such mismatches, as long as a reasonable number of text pairs can link *rotor* to ротор and *airflow* to поток.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No 248005. I'm particularly grateful to Richard Forsyth, who suggested the idea of using anchors, as well as to other members of the TTC project, and also to Reinhard Rapp for our joint work on term alignment.

References

Adafre, S. and de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 62–69, Trento.

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of LREC2004*, Lisbon.

Baroni, M. and Bernardini, S. (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.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

de Groc, C. (2011). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pages 497–498, Washington, DC.

Forsyth, R. and Sharoff, S. (2011). From crawled collections to comparable corpora: An approach based on automatic archetype identification. In *Proc. Corpus Linguistics Conference*, Birmingham.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proc. COLING'10*, Beijing, China.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 664—-671, Prague, Czech Republic.

Munteanu, D. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proc. of International Conference on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006*, Sydney.

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95, Portland, Oregon.

Rapp, R., Sharoff, S., and Babych, B. (2012). Identifying word translations from comparable documents without a seed lexicon. In *Proc. the Eighth Language Resources and Evaluation Conference, LREC 2012*, Istanbul.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proc. of the Comparing Corpora Workshop at ACL 2000*, pages 1–6, Hong Kong.

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409.

Sharoff, S. (2007). Classifying web corpora into domain and genre using automatic feature identification. In *Proc. of Web as Corpus Workshop*, Louvain-la-Neuve.

Sharoff, S. (2010). In the garden and in the jungle: Comparing genres in the BNC and Internet. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 149–166. Springer, Berlin/New York.

Su, F. and Babych, B. (2012). Development and application of a cross-language document comparability metric. In *Proc. the Eighth Language Resources and Evaluation Conference, LREC 2012*, Istanbul.