Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora

Serge Sharoff, Reinhard Rapp and Pierre Zweigenbaum

"That is not said right," said the Caterpillar.
"Not quite right, I'm afraid," said Alice, timidly:
"some of the words have got altered."
Lewis Carroll, Alice's Adventures in Wonderland

1 Data-driven Turn

The beginning of the 1990s marked a radical turn in various NLP applications towards using large collections of texts. For translation-related studies this implied the use of parallel corpora, i.e. authentic translations. Probably the first research group to explore this approach was the one at the IBM Watson Centre [11]. However, the use of parallel data predates the appearance of the computer, as evidenced from the Rosetta Stone, which contained the same text in three languages, thus providing the vital clue to deciphering the Egyptian hieroglyphs by Jean-François Champollion in 1822 [12]. It is interesting that more modern computational methods are still used for solving somewhat similar tasks [48].

For producing statistically reliable results the corpora need to be large, while the usual sources of large parallel corpora are public organisations producing a large

S. Sharoff (✉)
University of Leeds, West Yorkshire, United Kingdom
e-mail: s.sharoff@leeds.ac.uk

R. Rapp
University of Mainz, Mainz, Germany
e-mail: reinhardrapp@gmx.de

P. Zweigenbaum
LIMSI, CNRS and ERTIM, INALCO, Paris, France
e-mail: pz@limsi.fr

S. Sharoff et al. (eds.), *Building and Using Comparable Corpora*,
DOI: 10.1007/978-3-642-20128-8_1, © Springer-Verlag Berlin Heidelberg 2013

de KDE ist ein Projekt zur Entwicklung Freier Software. Hauptprodukt ist die Software Compilation bestehend aus KDE Workspace (auf deutsch: KDE Arbeitsumgebung; früher: K Desktop Environment), einer Arbeitsumgebung, sowie vielen Zusatzprogrammen für den täglichen Gebrauch, den KDE Applications. KDE Workspace ist vorrangig für Computer gedacht, auf denen ein Unix-ähnliches Betriebssystem läuft, wie zum Beispiel BSD, Linux oder Solaris.

en KDE (pronounced /keɪdi:'i:/) is a free software project based around its flagship product, a cross-platform desktop environment designed to run on Linux, FreeBSD, Windows and Mac OS X systems. The goal of the project is to provide basic desktop functions and applications for daily needs as well as tools and documentation for developers to write stand-alone applications for the system.

Fig. 1 Wikipedia articles on KDE in German and English

amount of translations, which are available in the public domain (usually because of the status of such organisations). Examples of corpora frequently used in NLP research are the Canadian Hansards [38], European Parliament proceedings [49], or the United Nations documents [23]. Such repositories are often the main resource for testing new tools and methods in Statistical Machine Translation.

However, reliance only on existing parallel texts leads to serious limitations, since the domains and genres of texts from such institutional repositories often do not match well the targets of NLP applications, e.g., the accuracy of statistical machine translation crucially depends on a good match between the training corpus and the texts to be translated [5, 22]. Also many more texts are produced monolingually in each language than produced by professional translators. This is the reason why many researchers have switched to using comparable (=less parallel) resources to mine information about possible translations. The importance of this research strand was first recognised in the 1990s [29, 64].

2 Collecting Comparable Resources

2.1 Degrees of Comparability

It is important to note that the distinction between comparable (non-parallel) and parallel corpora is not a clear-cut line. Informally any collection of texts covering two different languages can be measured along the scale of 'fully parallel' to 'non-related' with several options in between.

Parallel texts

These are traditional parallel texts, which can be classified into:

- Texts which are true and accurate translations, such as the UN or EuroParl documents;

- Texts which are reasonable translations with minor language-specific variations, e.g., an example of search in the OpenOffice user manuals for *New York* might be replaced with *Beijing* in the Chinese version;

Strongly comparable texts

They are heavily edited translations or independent, but closely related texts reporting the same event or describing the same subject. This category includes:

- Texts coming from the same source with the same editorial control, but written in different languages, e.g. the BBC News in English and Romanian [58];
- Independently written texts concerning the same subject, e.g. Wikipedia articles linked via iwiki, see Fig. 1 from Wikipedia, or news items concerning exactly the same specific event from different news agencies, such as AFP, DPA and Reuters;
- In exclusively oral languages, multiple recordings of a shared story [51]; once transcribed and augmented with an English gloss, they provide a comparable corpus in which correspondences can be searched;
- In sign languages, another instance of languages which do not come in the form of written texts, translations or multiple narrations of a same story: [73] outline how the gradation of parallel to comparable corpora can apply to sign language corpora in one or multiple languages.

Weakly comparable texts

This category includes:

- Texts in the same narrow subject domain and genre, but describing different events, e.g., parliamentary debates on health care from the Bundestag, the House of Commons and the Russian Duma;
- Texts within the same broader domain and genre, but varying in subdomains and specific genres, e.g., a crawl of discussion forums in information technology might bring more technical discussions on Linux server administration in English vs more user-oriented discussions on AutoCAD drawing issues in French.

Unrelated texts

This category comprises the vast majority of Internet texts, which can still be used for comparative linguistic research. For example, one can use random snapshots of the Web for Chinese, English, German and Russian to deliver comparable language teaching materials for these languages [47, 74].

2.2 Measuring Comparability

There is an inevitable trade-off between the amount of noise and the amount of data along this scale: fewer texts are translated than produced monolingually, fewer

events are covered by news agencies in exactly the same way in many languages than the number of monolingual stories in each of these languages. On the other hand, more parallel collections tend to be more useful for NLP applications, since more information can be extracted from greater parallelism in their content. In the 1990s and the beginning of the 2000s, work in computational translation studies (Statistical Machine Translation and Terminology Extraction) was mostly based on parallel corpora. Weakly-comparable and unrelated corpora have not been used a lot in computational research. Research presented in this volume (and in the events of the BUCC workshop series which preceded it) uses 'strongly-comparable' corpora.

In addition to these informal ways of assessing comparability, a more formal definition is based on measuring the distance between texts in their similarity space. This distance in the case of monolingual documents was first discussed by Adam Kilgarriff using the BNC as a benchmark: a *Known Similarity Corpus* was composed of documents known to be inherently similar within each category, while considerably different across the categories [46]. The distance between the documents in this approach can be measured by the degree of overlap between their keywords. There can be some difference in the way the keywords are extracted (top 500 words as used in [46], tf*idf, ll-score, etc.), as well as how the exact distance measure is defined (χ^2 in [46], cosine, Euclidean, etc.). Although, this suggestion from Kilgarriff was done within the same language, the idea can be extended further to measure corpus comparability by "translating" the documents from another language using either MT or simple dictionary mapping. Alternatively, instead of using more common words it is also possible to use Hapax Legomena (words occuring only once in each document) in order to identify potentially parallel documents [25, 59]. The advantage of this approach for closely related languages is that it makes it possible to by-pass the unreliable dictionaries and MT systems, while proper names and dates tend to be identical. If the aim is to investigate the relations between noisier collections (weakly comparable), it is possible to rely on classification of texts into topics and genres [75] under the assumption that the same labels are used for each language.

Irrespective of the approach to *measuring* text similarity, a benchmark for its *evaluation* is needed, which can be set in several ways:

- By using document-aligned parallel resources, such as Europarl or mining new parallel text collections;
- By using document-aligned comparable corpora, such as the dumps of the Wikipedia articles with information about their wiki categories and iwiki links between the languages;
- By collecting comparable resources using well-aligned keywords sent to a search engine, e.g., *autosave, configuring, debugger, user-friendly* for English versus *autoguardar, configurar, depurador, amigable* for Spanish [7].

Each of the approaches has its advantages and disadvantages. Evaluation using document-aligned parallel resources relies on texts which are known to be identical in terms of their topics, but such evaluation underestimates the degree of variation possible in comparable, originally produced texts. At the same time, a procedure

for collecting comparable documents from the Web needs its own evaluation on the accuracy of the collection procedure. One way out of this loop is by using extrinsic evaluation, i.e., by judging how suitable a comparable corpus is for a given multilingual task, such as extraction of parallel sentences or terms [8] or, better, a more finalized task such as cross-language information retrieval [53].

2.3 Monolingual Comparable Corpora

Comparable corpora are usually built by selecting two different languages, specifying a set of dimensions (topic, genre, time period, etc.) and selecting texts in these two languages with similar values for these dimensions.

However, monolingual comparable corpora can also be built. In this case the language dimension is fixed and it is one of the other dimensions which varies, for instance the intended audience (domain specialists versus lay people [19, 24]), the time period (e.g., nineteenth century press vs contemporary press [71]) or the source (different news agencies reporting on events in the same time period [83]).

Finding word or term 'translations' across these new varying dimensions presents different questions: it is facilitated by the proximity of language and large number of shared words in the two parts of the corpus. For instance, word overlap allows [83] to pair documents and then sentences; word alignment is then used to identify paraphrase segments. In [19], the dimension of variation is the intended audience: morphosemantic relations are used to detect matching expressions in lay and specialized corpora, reflecting differences in the patterns used in these two discourse types. These morphosemantic relations are discovered in [18] through POS-tagged n-gram alignment, taking into account linguistically-motivated morphosemantic variations.

2.4 Mining Parallel and Comparable Corpora

The easiest way of mining parallel corpora is by directly re-using the output of translation work in the form of segment-aligned TMX files, such as coming from TAUS Data Association.[1] The problem is that the number of texts available in this form is limited. More parallel texts are directly accessible in the form of multilingual webpages, such as newspapers or corporate websites. Earlier attempts at collecting such documents were based on the possibility to map structural similarities between the links to such websites, e.g., http://europa.eu/index_bg.htm vs http://europa.eu/index_el.htm which differ in the language identifier within their URLs [13, 69].

More modern approaches add the possibility of enriching the link heuristics with information about the contents [6]. Discovery of such websites can also be automated [26, 81]. Another possibility for getting good-quality parallel data is to mine parallel

[1] http://www.tausdata.org/

RSS feeds [80]. Such approaches can help in finding sufficient amounts of parallel texts even for medium-density languages, such as Hungarian [82] or Ukrainian [4].

Moving further down the cline towards comparable corpora, similar techniques can be used for extracting parallel texts from comparable websites by their structural links [2] or their contents [58, 79].

With respect to the collection of comparable resources using topical crawlers, there has been an array of recent EU projects, all aimed at designing tools for utilising bilingual information in crawling [8, 9, 77].

3 Using Comparable Corpora

3.1 Extraction of Bilingual Dictionaries

This section aims at exemplifying the wealth of work in comparable corpora by looking in some detail at one particular subtopic: Extracting information on word translations automatically from corpora (often referred to as *bilingual lexicon extraction*), rather than compiling dictionaries in the traditional lexicographic way, is an established application of parallel and comparable corpora.

With their seminal papers [10, 11], Brown et al. showed that information on word translations (the so-called translation models) could be reliably and in high quality extracted from parallel corpora, which was confirmed by others (e.g. [32]). But parallel corpora were (and, although to a lesser degree, still are) a scarce resource, so some years later the idea came up whether it might also be possible to derive information on word translations from comparable corpora. Independently of each other, at ACL 1995 Fung [27] and Rapp [64] suggested two approaches on how this could be accomplished. Fung [27] utilized a context heterogeneity measure, thereby assuming that words with productive context in one language translate to words with productive context in another language, and words with rigid context translate into words with rigid context. In contrast, the underlying assumption in Rapp [64] is that words which are translations of each other show similar co-occurrence patterns across languages. For example, if the words *teacher* and *school* co-occur more often than chance in English, then the same can be expected for their translations in a corpus of another language.

The validity of this co-occurrence constraint is obvious for parallel corpora, but it also holds for non-parallel corpora. It can be observed that this constraint works best with parallel corpora, second-best with comparable corpora, and somewhat worse with unrelated corpora. Robustness is not a big issue in any of these cases. In contrast, when applying sentence alignment algorithms to parallel corpora, omissions, insertions, and transpositions of text segments can have critical negative effects. However, the co-occurrence constraint when applied to comparable corpora is much weaker than the word-order constraint as used with parallel corpora. This is why larger corpora and well-chosen statistical methods are needed.

It should be noted that the advantages of looking at comparable rather than parallel corpora are not only robustness and ease of acquisition, but also that usually fewer corpora are required. Let us assume, for example, that we are interested in extracting dictionaries covering all possible pairs involving 10 languages, which would be altogether 90 directed language pairs. As both parallel and comparable corpora can be used in both directions of a language pair, this effectively reduces to 45 pairs. To deal with these 45 pairs, in the comparable case we need 10 corpora, one for each language. But in the parallel case we may need up to 45 corpora, thereby assuming that each language pair is based on the translation of a different text. That is, in the comparable case the required number of corpora increases linearly with the number of languages considered, but in the parallel case it can increase quadratically. However, if we are lucky, the same text may have been translated into several or all languages of interest. This means that the number of parallel corpora required can be reduced significantly. This is one of the reasons why large multilingual corpora covering many languages, such as Europarl and JRC-Acquis are particularly useful.

The task of identifying word translations has become one of the most investigated applications of comparable corpora. Following Rapp [64], most work was done using vector space approaches based on a multitude of variations of the above co-occurrence constraint (which can be seen as an extension of Harris' distributional hypothesis [40] to the multilingual case). Among the pioneers, Tanaka and Iwasaki [78] pursued a matrix-based approach where the selection of a target word candidate is seen in analogy to word sense disambiguation. Fung and McKeown [30] used word relation matrices which, using dictionary information, are mapped across languages to find new translation pairs. The accuracy is reported to be around 30 %. Fung and Yee [28] introduce an Information Retrieval inspired vector space approach: Using an existing bilingual lexicon of seed words, the co-occurrence vector of a word to be considered is translated into the target language. Then, using standard vector similarity measures, the resulting target language vector is compared to the vectors of the words in the target language vocabulary. The target language word with the highest similarity is considered to be the correct translation.

Peters and Picchi [61, 62] apply such a method for cross-language information retrieval. Given a query term to be translated, they compute its characteristic context words, and then translate these using existing dictionaries. They then search for those passages in the target language where there is a significant presence of the translated context words. This way, for any query term of interest, they obtain a ranked list of documents containing equivalent terms in another language.

Rapp [65] further refines the vector space approach, thereby also taking word order into account. This leads to an accuracy of 72 % for a standard test word list commonly used in Psychology. In subsequent work, transitivity across languages is taken into account [67]. Hereby advantage is taken of the possibility that, if corpora of more than two languages are available, the translations from one language to another can be determined not only directly, but also indirectly via a pivot language. This way, the more languages are considered the more evidence for a particular translation assignment can be provided by mutual cross-validation.

A related but different concept, referred to as *bridge languages*, had been used before by Schafer and Yarowsky [72]. However, the emphasis here is on cognate similarity between closely related languages such as Czech and Serbian. That is, if a Czech to English dictionary is available, the English translations of Serbian words can be determined by computing their orthographically most similar Czech counterparts, and by looking up their translations.

In addition to the bridge language concept, they manage to avoid the need for a seed lexicon by successfully combining temporal occurrence similarity across dates in news corpora, cross-language context similarity, weighted Levenstein string edit distance, and relative frequency and burstiness similarity measures.

A similar multi-clue approach is also used by Koehn and Knight [50]. They utilize spelling similarity, the above mentioned co-occurrence constraint, a second-order co-occurrence constraint (e.g. *Wednesday* and *Thursday* have similar contexts, as do their translations in another language), and corpus frequency (which should correlate between translations). They report a 39 % accuracy on a test set consisting of the 1,000 most frequent English and German nouns.

The potential of the spelling similarity clue is also demonstrated by Gamallo Otero and Garcia [33]. By extracting translation equivalents with similar spelling from Portuguese and Spanish comparable corpora (Wikipedia), they were able to come up with 27,000 new pairs of lemmas and multiwords not found in existing dictionaries, with about 92 % accuracy.

An additional potentially interesting clue which can be seen as an extension of spelling similarity is described in Langlais et al. [52]. In the medical domain they use analogical learning to exploit the formal similarity of medical words in some languages (systematic compounding). Their system does not require corpora but is trained on an initial bilingual lexicon.

Chiao and Zweigenbaum [15] conduct co-occurrence based lexicon extraction in the medical domain and systematically test several weighting factors and similarity measures. They found that by introducing an additional reverse-translation filtering step the accuracy of their system could be improved from 50 to 70 %. This is further elaborated in Chiao et al. [14].

Also specializing on the medical domain, for bilingual lexicon extraction Dejean et al. [17] not only exploit a seed lexicon but also a readily available multilingual medical thesaurus. They could show that using hierarchical information contained in the thesaurus significantly improves results.

Gamallo Otero and Pichel Campos [34] extract bilingual pairs of lexico-syntactic patterns from a parallel corpus. Subsequently they construct context vectors for all source and target language words by recording their frequency of occurrence in these patterns. There is thus only one vector space for both languages, so that vectors can be readily compared. For the language pair English–Spanish they report an accuracy of 89 % for high-frequency words. The method is further refined by Gamallo Otero and Pichel Campos in [35].

Shezaf and Rappoport [76] describe an algorithm introducing so-called non-aligned signatures for improving noisy dictionaries. The algorithm is in effect similar

to Fung and Yee [28] and Rapp [65], but (like [68]) rather than full co-occurrence vectors considers only salient context words (i.e. strong word associations).

As an application in contrastive linguistics, Defrancq [16] conducted a study for establishing cross-linguistic semantic relatedness between verbs in different languages based on monolingual corpora. A small number of verbs were semi-automatically investigated for their co-occurrences with particular interrogative elements, and then verbs were compared using Kullback-Leibler divergence.

Gaussier et al. [37], in an attempt to solve the problem of different word ambiguities in source and target language, use a geometric view and try to decompose the word vectors according to their senses. They investigate a number of methods, including canonical correlation analysis, multilingual probabilistic latent semantic analysis, thereby involving Fisher kernels. The best results with an improvement of 10 % are reported for a mixed method.

In contrast to the dominating vector space approaches based on word-co-occurrence data, Michelbacher et al. [21, 54] use linguistic relations like subcategorization, modification and coordination in a graph-based model. Also, other than most previous work, in their approach they distinguish between different parts of speech. Their basic approach is to use the SimRank algorithm to recursively compute node similarities. These are based on the similarity scores of neighboring nodes within a graph. Dorow et al. [21] proposed an extension towards cross-lingual semantic relatedness. It computes node-similarities between two graphs and allows for weighted graph edges.

Garera et al. [36] use a vector space model but consider dependency links rather than word co-occurrences. By doing so they obtain an improvement of 16 % for the language pair English-Spanish. They induce translation lexicons from comparable corpora based on multilingual dependency parses which takes long-range dependency into account. The system is shown to bring a 16 to 18 % improvement over a co-occurrence-based baseline. A similar approach is also pursued by Yu and Tsujii [84]. Their work is also based on the observation that a word and its translation share similar dependency relations, and they also obtain significant improvements.

There have also been a number of attempts to generate bilingual dictionaries from comparable corpora without the need of a seed lexicon. Diab and Finch [20] do so by using a computationally expensive bootstrapping approach which only requires very few seed translations. Otherwise their approach is related to Rapp [64], but they limit the co-occurrences they consider to those between the top 2,000 frequent tokens in the corpus and the top 150 frequent tokens, in four different collocation positions. Their method for searching new word translations is based on a gradient descent algorithm. They iteratively change the mapping of a given word until they reach a local minimum for the sum of squared differences between the association measure of all pairs of words in one language and the association measure of the pairs of translated words. Their reported accuracies are between 92.4 and 98.7 %, but for a pseudo translation task using two different corpora of the same language. So it might be a challenge to make the algorithm converge for non-related languages.

Haghighi et al. [39] approach the task of bilingual lexicon extraction by looking at word features such as co-occurrence counts and orthographic substrings, and

then inducing translations using a generative model based on canonical correlation analysis, which explains the monolingual lexicons in terms of latent matchings. For a range of corpus types and languages they show that high-precision lexicons can be learned even without a seed lexicon.

Robitaille et al. [70] deal with bilingual dictionary construction for multi-word terms. For their list of seed terms they download taylor-made multilingual corpora from the Web. They then extract multi-word terms from these corpora, and use a compositional method to align them across languages. Coverage is increased using a bootstrapping method.

The following three publications replace seed lexica by Wikipedia interlanguage links, which are pointers between wikipedia articles in different languages that relate to the same headword. Hassan and Mihalcea [41] represent words using explicit semantic analysis, and then compute the semantic relatedness of these concept vectors across languages by exploiting the mappings from the Wikipedia interlanguage links. Rapp et al. [66] do something similar but replace explicit semantic analysis by a keyword extraction procedure used for representing documents, and then applying an alignment algorithm on the keyword lists. Both methods show a reasonably good performance and can be applied to other multilingual document collections as well if these are aligned at the document level. (Such alignments can be computed using algorithms for measuring document comparability, which, however, usually require a bilingual lexicon). Prochasson and Fung [63] also start from aligned Wikipedia (and other) documents. They conduct a supervised classification and then utilize context-vector similarity and a co-occurrence model between words of aligned documents in a machine learning approach.

Morin and Prochasson [56] present an effective way of extracting bilingual lexica. By utilizing structural properties of the documents they extract parallel sentences from the comparable corpora, and then extract the dictionaries from these. Hazem and Morin [42] treat the dictionary extraction task as a question answering problem and describe their respective system QAlign. In a previous paper Morin et al. [55] showed that the quality of the comparable corpus is very important for dictionary construction.

The problem that most methods for dictionary extraction from comparable corpora have difficulties with rare words had been discovered early, but was for the first time put in focus by Pekar et al. [60]. Their solution was to estimate missing co-occurrence values based on similar words of the same language. Note, however, that the more recent approaches utilizing aligned comparable corpora [41, 63, 66] serve the same purpose and are likely to produce better results.

Finally, let us mention that, as shown by Rapp and Zock [68] bilingual lexica can even be extracted from monolingual corpora just by computing the strongest associations of foreign words occurring in a corpus. The reason is that in the contexts of foreign words often their translations are mentioned. But of course this is only of practical value for languages which are often cited, such as English. However, these can serve as pivots, thus mediating translations between other language pairs. In this method co-occurrence information is solely required for the target language. For the source language, to identify what counts as a foreign word, only a vocabulary list is

needed. Such a list can be extracted from a source language corpus, which relates the method to the comparable corpora topic.

3.2 Comparable Corpora for Statistical Machine Translation

Bilingual lexicons can be extracted with good success from parallel segments which have been extracted from comparable corpora. Given the limited availability of parallel corpora in many domains and for many language pairs, comparable corpora are often regarded as a potential source to help train Statistical Machine Translation (SMT) systems.

Most work in that area has been geared towards extracting parallel sub-parts of comparable corpora [44, 57, 85]. Using the collected parallel sub-parts helps train an SMT system and improve its performance over using a much larger out-of-domain parallel corpus. For instance, Abdul Rauf and Schwenk [1] obtained an increase of 2pt in BLEU score on Arabic to English translation, whereas Gahbiche-Braham et al. [31] increased by 6pt their BLEU score for Arabic to French translation.

In contrast, word translations directly extracted from comparable corpora currently have a too low precision to be useful for SMT. However they have been shown recently to improve the performance of a state-of-the-art cross-language information retrieval system [53], which indicates that further improvements in this line of research might pave the way to to applicability.

Another motivation for using comparable corpora in MT research can come from a cognitive perspective: Experience shows that persons who have learned a second language completely independently from their mother tongue can nevertheless translate between the languages. That is, human performance shows that there must be a way to bridge the gap between languages which does not rely on parallel data (in the context of human language learning with "parallel data" we could e.g. mean the use of mixed language in class). Using parallel data for MT is of course a nice shortcut and apparently much easier than understanding human language capabilities. But let us compare this approach to writing a chess program which simply enumerates very many possibilities of potential moves. This also tells us close to nothing about human reasoning. But language is not a domain as limited as chess. Therefore, in the long run it is likely that we will not get around understanding more about human language processing, and avoiding shortcuts by doing MT based on comparable corpora may well be a key to this.

4 Future Research Directions

The history of the BUCC workshops and the contributions to this volume identify several sources of interesting results. One comes from the fact that the Web is huge and it is getting easier to obtain reasonably similar texts for a range of languages.

The use of inter-wiki links in Wikipedia is a simple example of the growing space of similar texts. This presses the algorithms for more targeted detection of parallel and quasi-parallel segments in large collections, on the level of websites, documents, paragraphs and sentences. This leads to the possibility of using weakly comparable collections with the advantage of getting more closely related data for small domains (like wind energy) or less common language pairs (like German-Chinese).

Combination of features and supervision

Numerous types of information and functions on these types of information have been brought to bear to help identify matching words in comparable corpora: frequency of occurrence, co-occurrence counts, counts of lexico-syntactic patterns [34] or of dependency relations [3, 36], association measures, similarity measures, part-of-speech, cognates [72] and formal similarity [52], named entities and their relations [45], hierarchical information [17], co-presence in aligned comparable documents [63], to name but a few.

Most authors have contrasted these sources of information and tried to select those which worked best. Another path could be instead to try to combine them all together, pooling on the strengths of each type of information and function. This has been tried in only limited ways until now [43, 63]. Besides, most work has been performed in an unsupervised framework, whereas supervision is readily available in the standard setting through the availability of a partial bilingual dictionary. Supervision has proved very effective when used [3, 63]. Considering each type of information and each function on these types of information as features input to a supervised classifier might be a way to weight and combine them in an optimal way to identify word translations in comparable corpora, taking the best of each world.

References

1. Abdul Rauf, S., Schwenk, H.: Exploiting comparable corpora with TER and TERp. In: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora, pp. 46–54. Association for Computational Linguistics, Singapore (August 2009), http://www.aclweb.org/anthology/W/W09/W09-3109.pdf
2. Adafre, S., de Rijke, M.: 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), pp. 62–69. Trento (2006)
3. Andrade, D., Matsuzaki, T., Tsujii, J.: Learning the optimal use of dependency-parsing information for finding translations with comparable corpora. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pp. 10–18. Association for Computational Linguistics, Portland (June 2011), http://www.aclweb.org/anthology/W11-1203
4. Babych, B., Hartley, A., Sharoff, S.: Translating from under-resourced languages: comparing direct transfer against pivot translation. In: Proceedings of the MT Summit XI, pp. 412–418. Copenhagen (2007), http://corpus.leeds.ac.uk/serge/publications/2007-mt-summit.pdf
5. Babych, B., Hartley, A., Sharoff, S., Mudraya, O.: Assisting translators in indirect lexical transfer. In: Proceedings of 45^{th} ACL, pp. 739–746. Prague (2007), http://corpus.leeds.ac.uk/serge/publications/2007-ACL.pdf

6. Barbosa, L., Bangalore, S., Rangarajan Sridhar, V.K.: Crawling back and forth: using back and out links to locate bilingual sites. In: Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai (November 2011)
7. Baroni, M., Bernardini, S.: Bootcat: Bootstrapping corpora and terms from the Web. In: Proceedings of LREC2004. Lisbon (2004), http://sslmit.unibo.it/baroni/publications/lrec2004/bootcat_lrec_2004.pdf
8. Bel, N., Papavasiliou, V., Prokopidis, P., Toral, A., Arranz, V.: Mining and exploiting domain-specific corpora in the PANACEA platform. In: The 5th Workshop on Building and Using Comparable Corpora (2012)
9. Blancafort, H., Heid, U., Gornostay, T., Méchoulam, C., Daille, B., Sharoff, S.: User-centred views on terminology extraction tools: usage scenarios and integration into MT and CAT tools. In: Proceedings TRALOGY Conference "Translation Careers and Technologies: Convergence Points for the Future" (2011)
10. Brown, P., Pietra, S.D., Pietra, V.D., Mercer, R.: The mathematics of statistical machine translation: parameter estimation. Computat. Linguist. **19**(2), 263–312 (1993)
11. Brown, P.F., Cocke, J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computat. Linguist. **16**(2), 79–85 (1990)
12. Budge, E.A.T.W.: The Rosetta Stone. British Museum. London (1913)
13. Chen, J., Nie, J.: Parallel Web text mining for cross-language ir. In: Proceedings of RIAO, pp. 62–77 (2000)
14. Chiao, Y.C., Sta, J.D., Zweigenbaum, P.: A novel approach to improve word translations extraction from non-parallel, comparable corpora. In: Proceedings International Joint Conference on Natural Language Processing, Hainan (2004)
15. Chiao, Y.C., Zweigenbaum, P.: Looking for candidate translational equivalents in specialized, comparable corpora. In: COLING 2002 (2002)
16. Defrancq, B.: Establishing cross-linguistic semantic relatedness through monolingual corpora. Int. J. Corpus Linguist. **13**(4), 465–490 (2008)
17. Déjean, H., Gaussier, E., Sadat, F.: An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: COLING 2002 (2002)
18. Deléger, L., Cartoni, B., Zweigenbaum, P.: Paraphrase detection in monolingual specialized/lay comparable corpora. In: Sharoff, S., Rapp, R., Fung, P., Zweigenbaum, P. (eds.) Building and Using Comparable Corpora. Springer, Dordrecht (2012)
19. Deléger, L., Zweigenbaum, P.: Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In: Fung, P., Zweigenbaum, P., Rapp, R. (eds.) Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-Parallel Corpora, pp. 2–10. Association for Computational Linguistics, Singapore (August 2009), http://aclweb.org/anthology/W/W09/W09-3102
20. Diab, M., Finch, S.: A statistical wordlevel translation model for comparable corpora. In: Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO) (2000)
21. Dorow, B., Laws, F., Michelbacher, L., Scheible, C., Utt, J.: A graph-theoretic algorithm for automatic extension of translation lexicons. In: EACL 2009 Workshop on Geometrical Models of Natural Language Semantics (2009)
22. Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., Chen, Y.: Using moses to integrate multiple rule-based machine translation engines into a hybrid system. In: Proceedings of the Third Workshop on Statistical Machine Translation at ACL2008, pp. 179–182 (2008)
23. Eisele, A., Chen, Y.: MultiUN: A multilingual corpus from United Nations documents. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta (2010), http://www.euromatrixplus.net/multi-un/
24. Elhadad, N., Sutaria, K.: Mining a lexicon of technical terms and lay equivalents. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, pp. 49–56. Association for Computational Linguistics (2007)

25. Enright, J., Kondrak, G.: A fast method for parallel document identification. In: NAACL / Human Language Technologies, pp. 29–32. Rochester (2007)
26. Esplà-Gomis, M., Forcada, M.L.: Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. Prague Bull. Math. Linguist. **93**, 77–86 (2010)
27. Fung, P.: Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In: Proceedings of Third Annual Workshop on Very Large Corpora, pp. 173–183. Boston (1995)
28. Fung, P.: Extracting key terms from chinese and japanese texts. Int. J. Comput. Process. Orient. Lang. **12**(1), 99–121 (1998)
29. Fung, P.: A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In: Machine Translation and the Information Soup, pp. 1–17. Springer, Berlin (1998), http://www.springerlink.com/content/pqkpwpw32f5r74ev/
30. Fung, P., McKeown, K.: Finding terminology translations from non-parallel corpora. In: Proceedings of the 5th Annual Workshop on Very Large Corpora, pp. 192–202 (1997)
31. Gahbiche-Braham, S., Bonneau-Maynard, H., Yvon, F.: Two ways to use a noisy parallel news corpus for improving statistical machine translation. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pp. 44–51. Association for Computational Linguistics, Portland (June 2011), http://www.aclweb.org/anthology/W11-1207
32. Gale, W., Church, K.: A program for aligning sentences in bilingual corpora. Comput. linguist. **19**(1), 75–102 (1993)
33. Gamallo Otero, P., Garcia, M.: Extraction of bilingual cognates from wikipedia. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigo, F. (eds.) Computational Processing of the Portuguese Language. Lecture Notes in Artificial Intelligence, vol. 7243, pp. 63–72. Springer, Berlin (2012)
34. Gamallo Otero, P., Pichel Campos, J.R.: An approach to acquire word translations from non-parallel texts. In: EPIA, pp. 600–610 (2005)
35. Gamallo Otero, P., Pichel Campos, J.R.: Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. Comput. Linguist. Intell. Text Process. **6008**, 473–483 (2010)
36. Garera, N., Callison-Burch, C., Yarowsky, D.: Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In: CoNLL 09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, p. 129137. Morristown (2009)
37. Gaussier, E., Renders, J.M., Matveeva, I., Goutte, C., Djean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, p. 526533. Barcelona (2004)
38. Germann, U.: Aligned Hansards of the 36th Parliament of Canada (2001), http://www.isi.edu/natural-language/download/hansard/
39. Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D.: Learning bilingual lexicons from monolingual corpora. In: Proceedings of ACL-08: HLT, pp. 771–779. Columbus (2008)
40. Harris, Z.: Distributional structure. Word **10**(23), 146–162 (1954)
41. Hassan, S., Mihalcea, R.: Cross-lingual semantic relatedness using encyclopedic knowledge. In: EMNLP (2009)
42. Hazem, A., Morin, E.: Qalign: a new method for bilingual lexicon extraction from comparable corpora. Comput. Linguist. Intell. Text Process. **7182**, 83–96 (2012)
43. Hazem, A., Morin, E.: Extraction de lexiques bilingues partir de corpus comparables par combinaison de reprsentations contextuelles. In: Proceedings of the TALN 2013. ATALA, Les Sables d'Olonne (2013), in Press
44. Hewavitharana, S., Vogel, S.: Extracting parallel phrases from comparable data. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pp. 61–68. Association for Computational Linguistics, Portland (June 2011), http://www.aclweb.org/anthology/W11-1209
45. Ji, H.: Mining name translations from comparable corpora by creating bilingual information networks. In: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora:

from Parallel to Non-parallel Corpora, pp. 34–37. Association for Computational Linguistics, Singapore (August 2009), http://www.aclweb.org/anthology/W/W09/W09-3107
46. Kilgarriff, A.: Comparing corpora. Int. J. Corpus Linguist. **6**(1), 1–37 (2001)
47. Kilgarriff, A.: Comparable corpora within and across languages, word frequency lists and the kelly project. In: Proceedings of workshop on Building and Using Comparable Corpora at LREC, Malta (2010)
48. Knight, K., Megyesi, B., Schaefer, C.: The [copiale] cipher. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pp. 2–9. Portland (June 2011), http://www.aclweb.org/anthology/W11-1202
49. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Proceedings of MT Summit 2005 (2005), http://www.iccs.inf.ed.ac.uk/ pkoehn/publications/europarl-mtsummit05.pdf
50. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition, pp. 9–16 (2002)
51. Lahaussois, A., Guillaume, S.: A viewing and processing tool for the analysis of a comparable corpus of kiranti mythology. In: Proceedings of the 5th Workshop on Building and Using Comparable Corpora, pp. 33–41. ELDA, Istanbul (2012)
52. Langlais, P., Patry, A.: Translating unknown words by analogical learning. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 877–886 (2007)
53. Li, B.: Measuring and improving comparable corpus quality. Ph.D. thesis, Universit de Grenoble, Grenoble (June 2012)
54. Michelbacher, L., Laws, F., Dorow, B., Heid, U.,, Schütze, H.: Building a cross-lingual relatedness thesaurus using a graph similarity measure. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta (2010)
55. Morin, E., Daille, B., Takeuchi, K., Kageura, K.: Bilingual terminology mining—using brain, not brawn comparable corpora. In: Proceedings of the 45^{th} Annual Meeting of the Association for Computational Linguistics, pp. 664-671. Prague, Czech Republic (2007)
56. Morin, E., Prochasson, E.: Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In: BUCC2011 (2011)
57. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist. **31**(4), 477–504 (2005)
58. Munteanu, D., Marcu, D.: Extracting parallel sub-sentential fragments from non-parallel corpora. In: Proceedings of International Conference on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006. Sydney (2006)
59. Patry, A., Langlais, P.: 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, pp. 87–95. Portland (June 2011), http://www.aclweb.org/anthology/W11-1212
60. Pekar, V., Mitkov, R., Blagoev, D., Mulloni, A.: Finding translations for low-frequency words in comparable corpora. Mach. Transl. **20**(4), 247–266 (2006)
61. Peters, C., Picchi, E.: Using linguistic tools and resources in cross-language retrieval. In: Hull, D., Oard, D. (eds.) Cross-Language Text and Speech Retrieval Papers from the 1997 AAAI Spring Symposium, pp. 179–188. AAAI Press, San Francisco (1997)
62. Picchi, E., Peters, C.: Exploiting lexical resources and linguistic tools in cross-language information retrieval: the EuroSearch approach. In: First International Conference on Language Resources & Evaluation, pp. 865–872. Granada (1998)
63. Prochasson, E., Fung, P.: Rare word translation extraction from aligned comparable documents. In: Proceedings of ACL-HLT, Portland (2011)
64. Rapp, R.: Identifying word translations in non-parallel texts. In: Proceedings of the 33rd ACL, pp. 320–322. Cambridge (1995)
65. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th ACL, pp. 395–398. Maryland (1999)

66. Rapp, R., Sharoff, S., Babych, B.: Identifying word translations from comparable documents without a seed lexicon. In: Proceedings of the Eighth Language Resources and Evaluation Conference, LREC 2012. Istanbul (2012)
67. Rapp, R., Zock, M.: Automatic dictionary expansion using non-parallel corpora. In: Fink, A., Lausen, B., Ultsch, W.S.A. (eds.) Advances in Data Analysis, Data Handling and Business Intelligence. Proceedings of the 32nd Annual Meeting of the GfKl, 2008. Springer, Heidelberg (2010)
68. Rapp, R., Zock, M.: The noisier the better: identifying multilingual word translations using a single monolingual corpus. In: Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING. pp. 16–25. Beijing (2010)
69. Resnik, P., Smith, N.: The Web as a parallel corpus. Comput. Linguist. **29**(3), 349–380 (2003), http://www.umiacs.umd.edu/ resnik/strand/
70. Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., Utsuro, T.: Compiling French-Japanese terminologies from the Web. In: Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics, pp. 225–232. Trento (2006)
71. Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., Zweigenbaum, P.: Structured named entities in two distinct press corpora: contemporary broadcast news and old newspapers. In: Proceedings of the Sixth Linguistic Annotation Workshop, pp. 40–48. Association for Computational Linguistics, Jeju, Republic of Korea (July 2012), http://www.aclweb.org/anthology/W12-3606
72. Schafer, C., Yarowsky, D.: Inducing translation lexicons via diverse similarity measures and bridge languages. In: Proceedings of CoNLL (2002)
73. Segouat, J., Braffort, A.: Toward categorization of sign language corpora. In: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora, pp. 64–67. Association for Computational Linguistics, Singapore (August 2009), http://www.aclweb.org/anthology/W/W09/W09-3111
74. Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In: Baroni, M., Bernardini, S. (eds.) WaCky! Working Papers on the Web as Corpus. Gedit, Bologna (2006), http://wackybook.sslmit.unibo.it
75. Sharoff, S.: In the garden and in the jungle: comparing genres in the BNC and Internet. In: Mehler, A., Sharoff, S., Santini, M. (eds.) Genres on the Web: Computational Models and Empirical Studies, pp. 149–166. Springer, Berlin (2010)
76. Shezaf, D., Rappoport, A.: Bilingual lexicon generation using non-aligned signatures. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. p. 98107. Uppsala (2010)
77. Skadiņa, I., Vasiljevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D., Gornostay, T.: Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In: Proc. 3rd Workshop on Building and Using Comparable Corpora. Malta (2010).
78. Tanaka, K., Iwasaki, H.: Extraction of lexical translations from non-aligned corpora. In: Proceedings of the 16th conference on Computational linguistics (COLING96), vol. 2, pp. 580–585 (1996)
79. Tillmann, C.: A beam-search extraction algorithm for comparable data. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 225–228 (2009)
80. Tsvetkov, Y., Wintner, S.: Automatic acquisition of parallel corpora from websites with dynamic content. In: Proceedings of The Seventh International Conference on, Language Resources and Evaluation (LREC-2010) (2010)
81. Uszkoreit, J., Ponte, J.M., Popat, A.C., Dubiner, M.: Large scale parallel document mining for machine translation. In: COLING '10: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1101–1109 (2010)
82. Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., Tron, V.: Parallel corpora for medium density languages. In: N. Nicolov, K. Bontcheva, G.A., Mitkov, R. (eds.) Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05, pp. 247–258. Benjamins (2007), http://www.kornai.com/Papers/ranlp05parallel.pdf

83. Wang, R., Callison-Burch, C.: Paraphrase fragment extraction from monolingual comparable corpora. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pp. 52–60. Association for Computational Linguistics, Portland (June 2011), http://www.aclweb.org/anthology/W11-1208
84. Yu, K., Tsujii, J.: Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In: Proceedings of HLT-NAACL 2009, pp. 121–124. Boulder (2009)
85. Zhao, B., Vogel, S.: Adaptive parallel sentences mining from Web bilingual news collection. In: Proceeding of the 2002 IEEE International Conference on Data Mining (ICDM 2002) (2002)

http://www.springer.com/978-3-642-20127-1

Building and Using Comparable Corpora
Sharoff, S.; Rapp, R.; Zweigenbaum, P.; Fung, P. (Eds.)
2013, XII, 335 p. 70 illus., 14 illus. in color., Hardcover
ISBN: 978-3-642-20127-1