Multilingwis2 – Explore Your Parallel Corpus

Graën, Johannes ; Sandoz, Dominique ; Volk, Martin

Multilingwis^2 – Explore Your Parallel Corpus

Johannes Graën, Dominique Sandoz, Martin Volk
Institute of Computational Linguistics
University of Zurich
{graen|volk}@cl.uzh.ch, dominique.sandoz@uzh.ch

Abstract

We present Multilingwis^2, a web based search engine for exploration of word-aligned parallel and multiparallel corpora. Our application extends the search facilities by Clematide et al. (2016) and is designed to be easily employable on any parallel corpus comprising universal part-of-speech tags, lemmas and word alignments.

In addition to corpus exploration, it has proven useful for the assessment of word alignment quality. Loading the results of different alignment methods on the same corpus as different corpora into Multilingwis^2 alleviates their comparison.

1 Introduction

In (ibid.), we introduced *Multilingwis* (Multilingual Word Information System), our approach for exploring translation variants of multi-word units in multiparallel corpora. It relies on a part-of-speech tagged and word-aligned parallel corpus as source material, a *PostgreSQL* database for efficient retrieval (see Graën, Clematide, et al. 2016) and a standard web server equipped with *PHP* for the user interface. Our corpus data comes from CoStEP (Graën, Batinic, et al. 2014), which is a cleaner version of the Europarl corpus (Koehn 2005), and comprises 240 million tokens in English, German, French, Italian, and Spanish.

We, subsequently, received several requests regarding the portability of our retrieval engine and search interface to other corpora. Our decision to decouple Multilingwis from the particular data structure that our corpus had grown into and to release a version that can easily be adopted to other corpora coincided with the introduction of a proximity search operator (see Bartunov and Sigaev 2016, pp. 14–23) into PostgreSQL's full text search engine (PostgreSQL Global Development Group 2017). This led to the redesign of Multilingwis' search engine to allow for more complex searches by combining our queries with a full text search vector index.

In this paper, we describe the preparatory steps to produce the required corpus data, the functionality of Multilingwis^2 and the background of our search engine.

2 Corpus Preparation

We discriminate between content and function words and define content words to be either adjectives, adverbs, nouns or verbs, which we tell apart by means of universal part-of-speech tags (Petrov et al. 2012). Any corpus to be used with Multilingwis^2 thus requires these tags. They can be obtained directly using a tagger that produces universal tags or indirectly by mapping the language-specific tagsets to the universal one.

In addition to tagging, lemmatization is required by Multilingwis to provide a lemma-based search. The new version of our search engine is also capable to perform searches on word forms, but the resulting translation variants are always conflated to lemma sequences.

For our own corpus, we use the *TreeTagger* (Schmid 1994) for both, tagging and lemmatization and apply a subsequent lemma disambiguation algorithm similar to the one described in (Volk et al. 2016). This step reduces the amount of ambiguous lemmas, i.e. those for which the TreeTagger had seen more than one lemma during training, but some lemmas remain ambiguous. While they will not match any regular search query, they might appear in the list of translation variants, though.

Alongside those annotations, word alignments (see Tiedemann 2011, ch. 5) are crucial for Multilingwis. Any translation variant is derived from

[2] This is the second version, not a footnote number.

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 247–250,
Gothenburg, Sweden, 23-24 May 2017.

the list of tokens aligned with a particular search hit. Word alignment is usually preceded by sentence alignment as word alignment tools are typically not capable of aligning whole documents.[1] For our corpus data, we used *hunalign* (Varga et al. 2005) for sentence alignment, which can be provided with a dictionary for a particular language combination, or learn the dictionary from the parallel documents using a two-pass bootstrapping approach.

Word alignment tools such as *Giza++* (Och and Ney 2003) or *fast_align* (Dyer et al. 2013) produce unidirectional alignments which need to be symmetrized to obtain symmetric alignments. This requirement does not apply to the *Berkeley Aligner* (Liang et al. 2006) whose models are trained to produce symmetric alignments in the first place. Multilingwis expects word alignments to be symmetric. Independent of whether they are symmetric or not, union symmetrization is performed during corpus initialization, which has no effect on already symmetric alignments.

Additional attributes used by Multilingwis for visualization purposes are: white spaces that have been deleted during tokenization and any meta information related to a particular document in form of attribute value pairs. All this information is optional and will merely be visualized if available.

3 Functionality

Multilingwis' search strategy used to be simple: starting from a sequence of lemmas[2], all occurrences of those lemmas in the given order and with nothing in between them but (at most three) function words were selected and the translation variants calculated on this basis (see Clematide et al. 2016, sec. 3). We now extend the search to allow for any combination of search terms. The standard search mode conforms with what most search engines do: they find documents in which all of the given terms appear. In addition, a sequence of search terms enclosed in brackets is expected to occur consecutively without any intermediate token (phrasal search expressions).

For all searches, the user can choose whether the search is based on word forms or lemmas and if function words should be ignored. Having chosen lemma search and to ignore function words, a search where all search terms are enclosed in brackets will yield multi-word units.[3] A combination of phrasal and non-phrasal search expressions facilitates the search of multi-word expressions with flexible and fixed parts, e.g. German [in Frage] stellen *'to question'* finds "Ich möchte das in Frage stellen." *'I would like to question it.'* as well as "Keiner stellt das in Frage." *"Nobody questions it."* in our corpus, whereas in Frage stellen (without the phrasal restriction) will also yield sentences such as "Diese Frage stellt sich in der Tat." *"This question arises as a matter of fact."*.

Placeholders in phrasal search expressions provide means to express variable positions in multi-word expressions such as "to keep one's head above water". The search query [keep * head above water] will match "They use drug dealing, theft, and small-scale crime as means of keeping their heads above water." and "We have been trying to keep our heads above water for years.".

In case meta information has been provided, the attributes can serve as a filter. Europarl comprises the debates of the European Parliament, where speakers typically use their native language. The information, which language has originally been used is available in 82 % of the speaker contributions and is of great value for linguist, as we have learned in various occasions where we presented Multilingwis. By providing the original language as meta information, we enable the user to limit their search to a particular source language.

The user interface allows to select the search language. If none has been selected, Multilingwis evaluates which languages comprise the search terms as word forms or lemmas (depending on the search mode) and picks the one with the highest frequency averaged over all results. In our corpus, the search con *'with'* and calma *'rest'* (together *'at rest'* in both languages) will prefer Spanish over Italian since 'con' is much more frequent in Spanish and 'calma' shows approximately the same frequency in both languages. The third-ranked option is the combination of preposition 'con' with adjective 'calmo', which comprises 'calma' as word form. While search is performed

[1] Shorter sentences provide less opportunities for wrong alignment. That is why we split sentences when we come across a colon or semicolon.

[2] The user was allowed to enter any sequence of word forms, which was transformed into a sequence of lemmas by a finite-state conversion mechanism built on the corpus data.

[3] That is the only search mode in the first version of Multilingwis.

using the first-ranked option, the user can explicitly select the search language, which will perform a search based on the top-ranked option in that language.

4 Search Engine

Searches are performed by a PostgreSQL database, which not only provides fast retrieval but also performs the aggregation of individual search hits to distributions of translation variants in all languages efficiently. The import of corpora into the database is done by means of a single tabular-separated input file (similar to the CoNLL format but extended with columns for all the information specified in section 2). Parting from that import data, Multilingwis reconstructs the hierarchical structure of the corpus (documents, sentences, tokens), replaces columns involved in search (word forms, lemmas, meta information) by foreign key relationships with numerical identifiers, calculates full text search vectors on word forms and lemmas for both search modes (all tokens or content words only), and extracts and symmetrizes word alignments.

The last but most important step in preparation of the database is to index all attributes that will be used in retrieval. We create an inverted index on each text search vector, so that the index can be queried for the occurrence of all search terms (in a particular positional configuration if required by phrasal search expressions). All other attributes are indexed by standard B-tree indices. For the word alignment relation, we use a composite index as described in (Graën, Clematide, et al. 2016).

At search time, one of the inverted indices is scanned according to the search configuration and the matching tokens account for the search hits. With these hits as basis, the word alignment index is used to retrieve the tokens aligned to each of source tokens. The sequence of lemmas of those aligned tokens constitute the translation variants that are subsequently counted separately per language and build the statistics of translation variants shown in the user interface. The order of the aligned tokens makes a difference, i.e. the same set of lemmas in different orders makes for different translation variants. This is to distinguish expression like "human rights violations" and "violations of human rights".

After searching, the list of hits and aligned tokens can be inspected. The results are ordered by common shortness, i.e. shorter sentences in all languages come first.[4] The user may filter the result list for individual sets of translation variants in all languages. If there is no corpus example agreeing with the intersection of those filters, an empty list is shown.

5 Conclusions

We present Multilingwis^2, an exploration tool for parallel corpora based on word-alignment. Unlike the first version of Multilingwis, search is not limited to lemmas, and function words are not ignored per se.

Our own search engine is equipped with three different corpora: a seven-language corpus extracted from CoStEP (Graën, Batinic, et al. 2014) covering English, German, Finnish, French, Italian, Polish, and Spanish, the Text+Berg corpus (Göhring and Volk 2011) and the Bulletin corpus (Volk et al. 2016), and can be accessed at `https://pub.cl.uzh.ch/purl/multilingwis2`.

We also provide the source code and an extended installation manual at the same place. We offer Multilingwis^2 to anyone interested in using it on their own corpus.

Acknowledgments

This research was supported by the Swiss National Science Foundation under grant 105215_146781/1 through the project "SPARCLING – Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation".

[4] The more the sentences deviate in length, the more likely they will have alignment errors.

References

Bartunov, Oleg and Teodor Sigaev (2016). "FTS is DEAD ? – Long live FTS !" `https://www.slideshare.net/ArthurZakirov1/better-full-text-search-in-postgresql`. Accessed March 12th, 2017.

Clematide, Simon, Johannes Graën, and Martin Volk (2016). "Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora". In: *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseologia computacional y basada en corpus: perspectivas monolingües y multilingües*. Ed. by Gloria Corpas Pastor. Geneva: Tradulex, pp. 447–455.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith (2013). "A Simple, Fast, and Effective Reparameterization of IBM Model 2". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–649.

Göhring, Anne and Martin Volk (2011). "The Text+Berg Corpus An Alpine French-German Parallel Resource". In: *Traitement Automatique des Langues Naturelles*, p. 63.

Graën, Johannes, Dolores Batinic, and Martin Volk (2014). "Cleaning the Europarl Corpus for Linguistic Applications". In: *Proceedings of the Conference on Natural Language Processing*. (Hildesheim). Stiftung Universität Hildesheim, pp. 222–227.

Graën, Johannes, Simon Clematide, and Martin Volk (2016). "Efficient Exploration of Translation Variants in Large Multiparallel Corpora Using a Relational Database". In: *4th Workshop on Challenges in the Management of Large Corpora Workshop Programme*. Ed. by Piotr Bański, Marc Kupietz, Harald Lüngen, et al., pp. 20–23.

Koehn, Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *Machine Translation Summit*. (Phuket). Vol. 5, pp. 79–86.

Liang, Percy, Ben Taskar, and Dan Klein (2006). "Alignment by Agreement". In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 104–111.

Och, Franz Josef and Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models". In: *Computational linguistics* 29.1, pp. 19–51.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). "A Universal Part-of-Speech Tagset". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari et al. Istanbul: European Language Resources Association (ELRA).

PostgreSQL Global Development Group (2017). *PostgreSQL 9.6 Documentation – Chapter 12. Full Text Search*. `https://www.postgresql.org/docs/9.6/static/textsearch.html`. Accessed March 12th, 2017.

Schmid, Helmut (1994). "Probabilistic part-of-speech tagging using decision trees". In: *Proceedings of International Conference on New Methods in Natural Language Processing*. (Manchester). Vol. 12, pp. 44–49.

Tiedemann, Jörg (2011). *Bitext Alignment*. Vol. 4. Synthesis Lectures on Human Language Technologies 2. Morgan & Claypool.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy (2005). "Parallel corpora for medium density languages". In: *Proceedings of the Recent Advances in Natural Language Processing*. (Borovets), pp. 590–596.

Volk, Martin, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel (2016). "Building a Parallel Corpus on the World's Oldest Banking Magazine". In: *Proceedings of the Conference on Natural Language Processing*. (Bochum).