Multilingwis2 – Explore Your Parallel Corpus

Graēn, Johannes; Sandoz, Dominique; Volk, Martin

Abstract: We present Multilingwis2, 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 Multilingwis2 alleviates their comparison.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-137129

Originally published at:
Multilingwis\textsuperscript{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\textsuperscript{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\textsuperscript{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 \textit{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\textsuperscript{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\textsuperscript{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 \textit{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...
the list of tokens aligned with a particular search hit. Word alignment is usually preceded by sen-
tence alignment as word alignment tools are typ-
ically not capable of aligning whole documents.1 For our corpus data, we used hunalign (Varga et al.
2005) for sentence alignment, which can be pro-
vided with a dictionary for a particular language
combination, or learn the dictionary from the par-
allel documents using a two-pass bootstrapping ap-
proach.

Word alignment tools such as Giza++ (Och and
Ney 2003) or fast_align (Dyer et al. 2013) produce
unidirectional alignments which need to be sym-
metrized to obtain symmetric alignments. This re-
quirement 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 sym-
metric. Independent of whether they are symmetric
or not, union symmetrization is performed dur-
ing corpus initialization, which has no effect on al-
ready symmetric alignments.

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

3 Functionality

Multilingwis’ search strategy used to be simple:
starting from a sequence of lemmas2, all occur-
cences of those lemmas in the given order and with
nothing in between them but (at most three) func-
tion words were selected and the translation vari-
ants 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 stan-
dard search mode conforms with what most search
ingines do: they find documents in which all of
the given terms appear. In addition, a sequence of
search terms enclosed in brackets will yield multi-word units.3 A combi-
nation of phrasal and non-phrasal search expres-
sions facilitates the search of multi-word expres-
sions with flexible and fixed parts, e.g. Ger-
man [in Frage] stellen ‘to question’ finds “Ich
möchte das in Frage stellen.” ‘I would like to ques-
tion 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 contribu-
tions 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, Multiling-
wis evaluates which languages comprise the search
terms as word forms or lemmas (depending on the
search mode) and picks the one with the high-
est 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 approxi-
mately the same frequency in both languages. The
third-ranked option is the combination of preposi-
tion ‘con’ with adjective ‘calmo’, which comprises
‘calma’ as word form. While search is performed

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

2The 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.

3That is the only search mode in the first version of Mul-
tilingwis.
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. 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.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”.

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


