Exploring Properties of Intralingual and Interlingual Association Measures Visually

Johannes Graën, Christof Bless
Institute of Computational Linguistics
University of Zurich
graen@cl.uzh.ch, christof.bless@uzh.ch

Abstract

We present an interactive interface to explore the properties of intralingual and interlingual association measures. In conjunction, they can be employed for phraseme identification in word-aligned parallel corpora.

The customizable component we built to visualize individual results is capable of showing part-of-speech tags, syntactic dependency relations and word alignments next to the tokens of two corresponding sentences.

1 Introduction

In corpus linguistics, statistical association measures are used to empirically identify words that attract each other, i.e. that appear together in a corpus significantly more often than pure change would let us expect. Several association measures have been proposed and motivated in different ways (for an overview see Evert 2004, 2008). What they have in common is that they provide a scale that allows for ordering: from high association to no measurable association to negative association.[1]

Association measures can not only be applied to monolingual corpora, where they help identifying collocations, but also to interlingual relations, in our case word alignments in parallel corpora. In (Graën 2017), we exploit the fact that while some words in parallel texts are regular translations of each other, others are forced by idiomatic constraints. To find these constraints and thus identify phrasemes, we combine intralingual association measures on syntactical relations with interlingual association measures on word alignments.

This paper describes the necessary steps to prepare our corpus, which association measures we defined and how the results can visually be explored through our graphical interface.

2 Corpus Preparation

We extracted parallel texts from the *Corrected & Structured Europarl Corpus (CoStEP)* (Graën et al. 2014), which is a cleaner version of the *Europarl* corpus (Koehn 2005).

For tagging and lemmatization, we used the *TreeTagger* (Schmid 1994) with the language models available from the TreeTagger's web page. To increase tagging accuracy for words unknown to the language model, we extended the tagging lexicons, especially the German one, with lemmas and part-of-speech tags for frequent words. In addition, we used the word alignment information between all the languages (see below) to disambiguate lemmas for those tokens where the TreeTagger provided multiple lemmatization options. This approach is similar to the one described by Volk et al. (2016).

On the sentence segments identified (about 1.7 million per language), we performed pairwise sentence alignment with *hunalign* (Varga et al. 2005) and based on that word alignment with the *Berkeley Aligner* (Liang et al. 2006).[2] To increase alignment accuracy, we not only calculated the alignments on the word form of all tokens, but also on the lemmas of content words.[3] For the latter, we mapped the tagsets of the individual languages to the universal tagset defined by Petrov et al. (2012)

[1] It is worth mentioning that some association measures do not differentiate between high positive and high negative associations. Our application only uses those that make this difference.

[2] The Berkeley Aligner employs a symmetrically trained alignment model, whereas other word alignment tools such as *Giza++* (Och and Ney 2003) or *fastalign* (Dyer et al. 2013) require an additional symmetrization step for obtaining symmetrical alignments. Symmetric alignment in the first place is to be preferred over symmetrization of two asymmetric alignments (cf. Tiedemann 2011).

[3] Here, we used the word form instead if no lemma was provided.

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 314–317,
Gothenburg, Sweden, 23-24 May 2017.

and defined content words to be those tokens that are tagged as either nouns, verbs, adjectives or adverbs.

We used the *MaltParser* (Nivre et al. 2006) to derive syntactical dependency relations in German, English and Italian. As there was no pretrained model available for Italian, we built one based on the *Italian Stanford Dependency Treebank (ISDT)*.[4] Each parser model uses particular language-specific dependency relations. Universal dependency relations (McDonald et al. 2013) could facilitate the definition of syntactic relations. For our purpose however (see next section), it suffices to identify the direct object relationship of verbs. Moreover, at the time we prepared the corpus, there were no ready-to-use universal dependency parsers available for the languages required. Mapping language-specific parser models to universal dependency relations is not as straightforward as mapping individual tagsets to universal part-of-speech tags (cf. Marneffe et al. 2014).

3 Interlingual Association Measures

We aim at identifying phrasemes, i.e. highly idiomatic multi-word units. In (Graën 2017), we employ the example of support verb constructions consisting of a verb and its direct object, where the verb "supports" the semantics of the expression leaving aside its own. A *walk*, for instance, cannot literally be *taken* or *given* (Spanish: *dar un paseo*, literally *'give a walk'*). Supporting verbs often show a "light" character, hence the alias light verb construction.

Following the example of support verb constructions, we regard all verbs with aligned direct objects in parallel sentences as candidates. There are four relations that can be evaluated: Besides the intralinguistic association measure on the verb and its direct object in each language, we can also measure the association of both verbs and both objects by using the same association measures on the interlinguistic relation of word alignment. While an intralinguistic association measure makes a statement about the relative frequency of two words appearing in a particular constellation in a monolingual corpus, an interlinguistic association measures makes a statement about the relative frequency of two words being aligned in a parallel corpus.[5]

In the – frequent – case that the supporting verbs are otherwise not common translations of each other, i.e. they show little attraction apart from the constellation we are looking at, the interlinguistic association measure of the aligned verbs yields a comparably low score. We exploit this fact and rank all candidates in such a way that for a high rank this verb alignment score is required to be low while all other scores are required to be high.

4 Visual Exploration

Different properties of the well-known association scores make them suited for different tasks and different levels of cooccurrence (Evert 2008, cps. 3–5). It is less clear though, what the characteristics of association on word alignment are. We, therefore, implemented an interface (depicted in Fig. 1) to explore the results of different association measures applied to particular patterns, which are described by syntactic relations and attributes.

This pattern is searched in the corpus, results are aggregated using the lemmas of all tokens and sorted by frequency of such lemma combinations.[6] The user can change the sort criterion to any of the following association measures: t-score, z-score, mutual information (MI), observed/expected ratio (O/E) and O^2/E. Both lemmas can be filtered using regular expressions.

When the user select a combination from the resulting list, the distribution of aligned lemma combinations of all available languages[7] is shown. The same association measures can be employed for sorting.

To explore the individual examples, we set up a visualization that displays the sentence and its aligned counterparts on top of each other. The user can navigate through all the sentences that include the selected linguistic constellation (source and target lemma combinations).

Two kinds of relations can be switched on and off: syntactic dependency relations between the words in both languages and the word alignments. In addition, universal part-of-speech tags are shown if requested. The tokens belonging to one of the lemma combinations are highlighted by

[4] `http://medialab.di.unipi.it/wiki/ISDT`

[5] When calculating association measures, we only take lemmas into account to reduce variation and get more reliable values.

[6] We do this for English, German and Italian as source languages and store the precalculated association measures in a database.

[7] Our corpus comprises alignments between English, Finnish, French, German, Italian, Polish and Spanish.

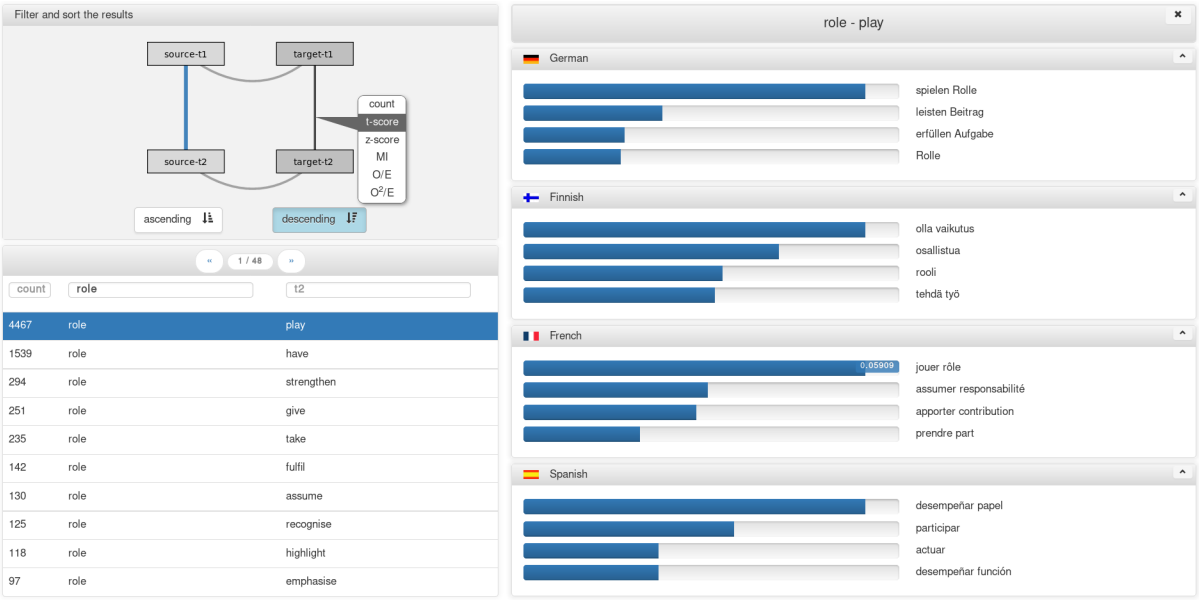

Figure 1: The support verb construction "play a role" and its translation into four other languages. Results in the target languages are sorted using t-score as association measure.

default. The attributes of all other tokens can be made visible interactively or switched on permanently. Integrating all this information on one page facilitates tracing differences in the usage of a particular linguistic schema.[8]

The graphical display is designed to be customizable and reusable. Its output can not only be used interactively, it also serves for printing as the graphics is rendered in Scalable Vector Graphics (SVG) format. Furthermore, the user can adjust the spacing between individual tokens and the gap between both sentences, and reposition dependency labels to achieve the best visual appearance.

5 Conclusions

We built an interface for exploration of different types of association measures. Intralingual association measures are widely used to assess the attraction of pairs of words in a corpus. Interlingual association measures do essentially the same but on word-alignments between corresponding sentences in two languages.

Our interface is an approach to visually explore the properties of different association measures. Results of a particular pattern applied to a source language and the aligned patterns in different target languages can be sorted according to a selection of association measures. Unlike the approach described in (Graën 2017), we do not (yet) provide the option of a weighted combined score.

The interface described here is available for exploring at: `http://pub.cl.uzh.ch/purl/visual_association_measures`. We also provide the source code of the visualization component there.

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".

[8] It also helps to detect recurring tagging, parsing or alignment errors.

References

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.

Evert, Stefan (2004). "The Statistics of Word Cooccurrences: Word Pairs and Collocations". PhD thesis. Universität Stuttgart.

– (2008). "Corpora and collocations". In: *Corpus linguistics: An international handbook* 2. Ed. by A. Lüdeling and M. Kytö, pp. 1212–1248.

Graën, Johannes (2017). "Identifying Phrasemes via Interlingual Association Measures". In: *Lexemkombinationen und typisierte Rede im mehrsprachigen Kontext*. Ed. by Christine Konecny et al. Tübingen: Stauffenburg Linguistik.

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.

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.

Marneffe, Marie-Catherine de, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). "Universal Stanford Dependencies: A cross-linguistic typology". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari et al. Vol. 14. European Language Resources Association (ELRA), pp. 4585–4592.

McDonald, Ryan T., Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee (2013). "Universal Dependency Annotation for Multilingual Parsing". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 2, pp. 92–97.

Nivre, Joakim, Johan Hall, and Jens Nilsson (2006). "Maltparser: A data-driven parser-generator for dependency parsing". In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Vol. 6, pp. 2216–2219.

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).

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).