Word order typology through interlingua word alignment

Robert Östling

Department of Linguistics, Stockholm University
SE-106 91 Stockholm
robert@ling.su.se

Abstract

With massively parallel coprora of hundreds or thousands of translations, it is possible to automatically perform typological studies of language structure using very large language samples. I investigate the domain of word order using multilingual word alignment and annotation transfer of a corpus with 1144 translations of the New Testament. Results are encouraging, with 85% to 95% agreement between the automatic system and the manually created World Atlas of Language Structures (WALS) for a range of different word order features.

1. Introduction

I have previously studied methods for multilingual word alignment of massively parallel corpora with over a thousand translations (Östling, 2014). In essence, the method developed uses Gibbs sampling in a Bayesian model to learn two things: a common "interlingua" representation of the text, and alignments from this representation to all of the individual translations.

One question left unanswered in the previous study was *why* one would want such a multilingual word alignment. There has been previous research on using massively parallel texts for investigations in linguistic typology, see for instance the special issue introduced by Cysouw and Wälchli (2007). Here, I present another application to linguistic typology: investigating word order features.

2. Method

The first step is to compute an interlingua alignment of the corpus, as described in my earlier work (Östling, 2014). Here, I use the same New Testament corpus, with 1144 translations in 986 different languages (some languages having multiple translations).

Second, the ten English translations are part-of-speech (PoS) tagged using the Stanford Tagger (Toutanova et al., 2003), converted to the Universal Part-of-Speech Tagset of Petrov et al. (2012), and dependency parsed with MaltParser (Nivre et al., 2007) trained on the Universal Dependency Treebank (McDonald et al., 2013) using MaltOptimizer (Ballesteros and Nivre, 2012).

Third, PoS and dependency annotations were transfered to the interlingua representation through direct multi-source projection. Given the fact that I use an alignment model based on the simplistic IBM Model 1, on a relatively short text (less than 8000 verses), a high amount of alignment errors is to be expected. Therefore, a very aggressive filtering scheme was used: only dependency links which are projected from at least 75% of source texts were included. In this way, alignment errors, divergent translations and sentences that are difficult to parse are excluded. This severely limits recall, but is acceptable since even a few tens of examples of each grammatical relation are usually sufficient to tell which ordering is dominant in a particular language.

Some experiments were performed using both German and English, with similar but somewhat worse results, possibly due to the fact that the interlingua was initialized with an English translation, and so is somewhat more easily alignable with English than with German.

Given the information available at this point, it is simple to compute which ordering of words in e.g. a verb-object relation is most frequent in a language. If multiple translations exist for a language, counts are aggregated per language in order to compare to WALS. Of course, comparing different translations in the same language could be an interesting project as well.

3. Experiments

WALS, the World Atlas of Language Structures (Dryer and Haspelmath, 2013), contains classifications of languages according to a large number of structural features. I will focus on five of these, summarized in table 1 along with the agreement between the algorithm and WALS, for the subset of languages that are present both in the relevant WALS chapter and the New Testament corpus. Languages where WALS gives an option other than one of the possible permutations (e.g. that a language does not have adpositions, or that there is no dominant verb-object order) are excluded from the counts.

3.1 Results

First of all, we can see that the agreement between the algorithm's output and the hand-classified WALS entries is high, in all cases much higher than with either the chance or the most-common-category baselines. The lowest agreement is obtained for chapter 81A (verb/subject/object), which expected since there are six possible permutations, as opposed to two for the other features.

It is reasonable to expect that languages more dissimilar to English, and therefore more difficult to transfer English annotation to, would obtain less reliable results. Possibly for this reason, agreement seems to be lower for uncommon word orders, such as object-subject-verb (OSV), although there are too few examples of these to draw any solid conclusions.

WALS	Agreement	N	Description
81A	85.7%	342	order of verb, subject and object
82A	90.4%	376	order of verb and subject
83A	96.4%	387	order of verb and object
85A	95.1%	329	order of adposition and noun (pre-/postposition)
87A	88.0%	334	order of adjective and noun

Table 1: Agreement between the algorithm and WALS. N is the number of languages that are both in the relevant WALS chapter and in the New Testament corpus. All features are binary except 81A, which can take six values.

Nevertheless, the strong results in spite of a large and diverse sample of languages indicate that the approach is feasible for exploratory large-scale word order investigations. In addition, the output contains not only a hard classification into word order types, but also a measure of how strong this tendency is and which alternative word orders are also common. There is no easy way of automatically evaluating this aspect of the data, but Bernhard Wälchli (p.c.) informs me that results look reasonable for a manually evaluated set of languages.

3.2 HMM alignment

I have repeated the experiment using an extension of the basic alignment model (Östling, 2014) with a Hidden Markov Model (HMM) distortion model, akin to Vogel et al. (1996). The alignments are, as expected, of much higher quality than the original Model 1-based algorithm when evaluated on the translations with Strong's numbers (Östling, 2014, section 4.2). Surprisingly, there was a much greater *disagreement* with WALS.

The reason for this counterintuitive result seems to be that the HMM-based alignments contain a bias towards the English word order, which results in the English feature values (subject-verb-object, adjective-noun, prepositions) incorrectly being predicted for many languages.

4. Future directions

There are many other structural properties of languages that could be investigated with high-precision annotation transfer in massively parallel corpora, not just regarding word-order but also within in domains such as negation, comparison and tense/aspect. While there are limits to the quality and types of answers obtainable, the main advantages of the kind of method presented here is that it provides quick, quantitative answers capable of guiding more thorough typological research.

On the technical side, there are various ways to extend the basic alignment algorithm, such as adding fertility parameters or using symmetrization methods. These may be able to improve accuracy, although section 3.2 suggests that this is by no means certain. It would also be useful for many typological investigations to align at the morpheme level, rather than the word level.

References

Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: An optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 58–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF - Language Typology and Universals*, 60(2):95–99.

Matthew S. Dryer and Martin Haspelmath. 2013. The world atlas of language structures online. `http://wals.info`.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith 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 (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135, 6.

Robert Östling. 2014. Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 123–127, Gothenburg, Sweden, April. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.