A dataset and metric to evaluate lexical extraction from parallel corpora

Barend Beekhuizen
University of Toronto, Mississauga
Department of Language Studies / Department of Linguistics
barend.beekhuizen@utoronto.ca

1 Introduction and motivation[1]

Over the past decades, many methods for the study of cross-linguistic variation in word meanings have been developed, such as the use of non-linguistic stimuli (Bowerman and Pederson, 1992), the elicitation of explications in a restricted set of primitives (Goddard and Wierzbicka, 2013), and the use of secondary data (dictionaries) compiled into databases like IDS (Key and Comrie, 2015) and CLICS (List et al., 2018). A recent method is the use of massively parallel corpora: corpora containing the same source text translated into many target languages. Such corpora allow the researcher to study how the same lexical item was translated into multiple target languages (Wälchli and Cysouw, 2012; Wälchli, 2014). The use of parallel corpora presents a unique lens on cross-linguistic variation by comparing usages of words across languages as they occur in natural discourse contexts.

The wider application of this approach would benefit from computational methods that can automatically extract translation equivalent expressions and determine which translation equivalent expressions belong to the same lemma, which requires high-quality word alignment (Jurafsky and Martin, 2009, ch. 25) and unsupervised morphological segmentation (Goldsmith, 2001), both open questions in Computational Linguistics. Currently, only one such procedure exists: The L-model of (Wälchli, 2014) starts with a seed set of lemmas in one language, using its pattern of co-occurrence with potential translation equivalents directly, and furthermore integrates a stemming function. A similar, but not fully automated method is *SuperPivot* (Asgari and Schütze, 2017): again starting with a seed set of lemmas in one language, the most strongly aligned words across all languages are extracted, which in turn help identify translation-equivalent sublexical strings, but not assign them to individual tokens.

2 Current project

Given the value of such models for lexical semantic typology, further model development would be desirable and the availability of data and metrics for the evaluation of such models would stimulate such development. This project aims to fill that gap. As data, we use translations of the Bible, a much-translated collection of religious texts. The only publicly available corpus was a set of 100 Bible translations (Christodouloupoulos and Steedman, 2015), but recent initiatives containing greater numbers of translations exist (McCarthy et al., 2020).

Given the interest in broad lexical coverage and using English as the pivot, we focused on 3 parts of speech (PoS): nouns, adjectives, and verbs, and lemmas from three frequency bins (all lemmas per million): 30-100 (*low*); 100-300 (*mid*), and 300+ (*high*). (The data was lemmatized and PoS tagged with SpaCy; Honnibal and Montani, 2017.) For each PoS-frequency combination, 10 words were sampled, making up 99 lemmas in total.[2] For each lemma, 30 tokens in English were sampled.

[1]Thanks, foremost to the annotators, without whom this project would have been impossible: Chahla Ben-Ammar, Maya Blumenthal, Crystal Chow, Hinako Fujiwara, Sadaf Kalami, Juliet Miinalainen, Mia Misic, Chelsea Saguil, and Sabrina Yu. I would furthermore like to acknowledge the support of a Connaught New Researcher Award and NSERC Discovery grant (RGPIN-2019-06917), and thank Bernhard Wälchli for sharing and explaining his code.

[2]Manually adding 9 further lemmas. The set consists of: **N-high:** bread, daughter, day, foot, hand, head, heart, land, law, mountain, priest, prophet, sea, son, thing, time; **N-mid:** book, darkness, door, elder, feast, generation, heaven, night, ruler, temple; **N-low:** authority, prayer, promise, scribe, self, table, tax, will, witness, woe; **A-high:** dead, evil, first, full, good, holy, mighty, old, other, right, same, young; **A-mid:** able, afraid, certain, chief, new, poor, righteous, unclean, whole, wise; **A-low:** eternal, faithful, false, free, last, loud, rich, sick, true, weak; **V-high** burn, fall, find, gather, kill, live, love, pass, put, take, turn; **V-mid:** carry, cast, get, hide, hold, open, prepare, receive, save, work; **V-low:** allow, beat, continue, desire, endure, glorify, justify, need, suffer, touch.

Target languages were selected for which native, heritage or advanced bilingual speakers could be recruited as annotators from the local community of linguistics students. The languages were Arabic, Croatian, Farsi, Finnish, French, Hebrew, Japanese, Mandarin Chinese, Serbian, and Tagalog. While containing exclusively Eurasian languages as well as two languages that are closely related varieties, the data set has 6 different language families, which forms a reasonably diverse starting point for model evaluation. Future efforts to annotate languages for which speakers are hard to find are planned. The data will be released upon completion.

3 Exploring the data

Annotation statistics: Three data statistics are of interest. First, each English lemma has on average 4.19 unique translation equivalent expressions across its 30 tokens. Verbs have more unique translations (6.17) than adjectives (3.71) and nouns (2.81). Second, we find significant between-language variation in the proportion of zero-annotation tokens (5% in Tagalog vs. 17% in Farsi). Finally, we find between-language variation in the proportion of the annotation tokens contain multiple elements (ranging from 7% in French to 40% in Hebrew (due to root-transfix splits) and 42% in Chinese (due to compounding).

Model evaluation: To evaluate the extraction models, we first define a metric to assess their quality. The metric would have to work despite label space mismatches between the annotated and extracted data (the annotator may have given an infinitival verb form, while the extracted form is a root). This suggests evaluation techniques from unsupervised clustering, such as the Rand Adjusted Index (Hubert and Arabie, 1985). However, that metric presupposes that every token is assigned to a single cluster, whereas annotations frequently consisted of zero or more than one expression. To accommodate this, another metric was designed, starting from the the intuition that lexical boundaries are a critical property of the data for lexical semantic typology. Given a series of extracted translation equivalent expressions $E = [e_1, \ldots, e_n]$ of length n, where each element e is a (potentially empty) set of expressions, and a corresponding series of annotated expressions $A = [a_1, \ldots, a_n]$, the error $\mathcal{E}(E, A)$ is given as:

$$\mathcal{E}(E, A) = \frac{1}{\binom{n}{k}} \sum_{i,j \in C(n,2)} \epsilon(i, j). \qquad (1)$$

model	ar	fa	fi	hr	ja	**tot.**
fa +u	.71	.57	.53	.60	.42	.57
fa +i	.70	.58	.51	. .61	.39	.56
mf + fa +u	.54	.43	.47	.47	.34	.45
mf + fa +i	.52	.39	.42	.39	.33	.41
L-model	**.40**	**.29**	**.41**	**.30**	**.32**	**.34**

Table 1: $\mathcal{E}$ per language, aggregated over lemmas.

That is: for each combination i, j of indices of E and A, the mean of all pairwise errors $\epsilon(i, j)$ is computed. If none of e_i, e_j, a_i, a_j is an empty set, $\epsilon(i, j)$ is the absolute difference between the Jaccard similarity of e_i and e_j and that of a_i and a_j (punishing the model when a pair of items in E and the corresponding pair in A are not equally similar). Otherwise, if e_i and a_i are both empty sets or both non-empty sets, and e_j and a_j are both empty or both non-empty sets, $\epsilon(i, j) = 0$. If the pairs e_i, e_j and a_i, a_j have mismatching empty sets, it counts as a full error on the model's part, and $\epsilon(i, j) = 1$. $\mathcal{E}$ ranges from 0 (absence of error) to 1 (the boundaries in E are orthogonal to those in A).

Models: Wälchli's L-model was tested here as the only published full extraction procedure. Readers are referred to the paper for fuller details. Informed baseline models were designed for comparison. All baseline models used word alignments from FastAlign (fa; Dyer et al., 2013), either with union (+u) or intersection (+i) symmetrization, and both settings applied either to the data in the target language as is, or as preprocessed with Morfessor 2.0 (mf; Virpioja et al., 2013). A 50% development/test split was applied to both languages and lemmas, and only results on the development lemmas for the developemnt languages are reported.

Results: Table 1 shows that the L-model performs best on all languages. Among the baseline models, morphological preprocessing helps in recognizing identical translation equivalents, while using the stricter symmetrization procedure seems to weed out incorrectly aligned candidates.

Despite its good performance, the L-model still displays avoidable errors. Future work would benefit from attending to the errors frequently found for the L-model: (too greedily) extracting cases where no or a less frequent expression is annotated, extracting a single element where multiple ones were annotated, and morphological missegmentations.

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124.

Melissa Bowerman and Erik Pederson. 1992. Topological relations picture series. In *Space stimuli kit 1.2*, page 51. Max Planck Institute for Psycholinguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. In *Proceedings of the 7th Language Resources and Evaluation Conference*, pages 375–395. Springer.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Cliff Goddard and Anna Wierzbicka. 2013. *Words and meanings: Lexical semantics across domains, languages, and cultures*. OUP Oxford.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Matthew Honnibal and Ines Montani. 2017. SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Daniel Jurafsky and James H Martin. 2009. *Speech and Language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson/Prentice Hall.

Mary Ritchie Key and Bernard Comrie, editors. 2015. *Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Simon J Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. Clics2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.

Arya D McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Berhard Wälchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50(3):671–710.

Bernhard Wälchli. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. In Benedict Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*, pages 355–393. Walter de Gruyter.