Multilingual generative models for selectional preference learning

An MPhil project proposal

A. Scoică (*as2270*), Girton College
November 30, 2012

Project Supervisor: Prof Stephen Clark & Dr Diarmuid Ó Séaghdha

Abstract

Selectional preference is the linguistic phenomenon that models the affinities or restrictions verbs may have on the type of arguments they can take in a given language model.

We hypothesize that selectional preferences are a direct consequence of world facts, and thus should be preserved across natural languages. However, to date there haven't been attempts at automatic selectional preference acquisition from more than two languages at once. We believe that coming up with a multilingual model capable of doing that will facilitate the transfer of knowledge between languages to improve results in resource-poor languages and possibly also improve accuracy in languages where plenty of resources are already available.

The project will build on previous work by extending bilingual vector alignment methods to more than two languages and by employing topic models for selectional preference learning. My proposed approach is to generalize the previous cross-language plausibility transfer model and to explore alternatives for its components.

1 Introduction

Current approaches to modelling selectional preferences fall into two main categories: class-based approaches and non-class approaches. The former try to map arguments onto a predefined class taxonomy, which is an expensive resource that needs to be designed separately. The latter methods avoid this need by either computing similarity measures without relying on an ontology, or by automatically inducing the set of classes. Examples of non-class approaches are: similarity based methods, which use distributional measures, discriminative models which employ automatic classification learned through both positive and negative examples, and generative probabilistic models which model predicate arguments as being generated by a latent variable.

Generative models treat predicate arguments as having been generated by a latent variable. Ó Séaghdha (2010) demonstrated that generative topic models such as Latent Dirichlet Allocation can be adapted to induce monolingual probabilistic models of selectional preference in the absence of a comprehensive taxonomy. However, since previous work on selectional preference induction has focused mainly on languages with large annotated corpora, such as English, these monolingual methods cannot be applied to cases of languages where rich resources are not available.

Still, starting from the assumption that selectional preferences are a consequence of world facts and do not depend arbitrarily on the language they are encoded in, it should be possible to transfer knowledge about them across languages without major losses.

Peirsman and Pado (2010) have already shown that cross-lingual knowledge transfer can be done with pairs of languages, where the plausibility of predicate-argument relations in a resource-poor source language can be approximated by the plausibility of a translation to a more resource-rich target language. They also proved that the knowledge transfer can take place even upon using unrelated corpora lacking lexical annotations in the two languages, because their proposed translation method relies on bootstrapping a bilingual vector space from the unparsed available corpora. However, they demonstrated that further performance gains are possible with only small amounts of annotated resources.

We believe that there is room for improvement in the model proposed by the authors and we have identified aspects where alternatives can be more investigated in more depth, such as: building a better aligned multilingual vector space using Wikipedia, extending the model to more than two languages, experimenting with various cross-language plausibility approximation formulae, experimenting with alternative translation procedures and improved smoothing using various types of similarity metrics.

Finally, given that Boyd-Graber and Blei (2009) proposed a framework capable of aligning topics between more than two languages using non-parallel corpora, and that Ó Séaghdha's results confirm the applicability of topic models to selectional preference learning, we also believe that the development of a multilingual generative model for selectional preference learning should be possible.

2 Approach & Outcomes

Initially, we will organize corpus data and automate the evaluation procedure. For languages included in the previous work by Peirsman and Pado, we will use the same datasets for a controlled comparison, while for new languages we will use similar data, the plausibility of which will be manually annotated.

We will first confirm the applicability of selectional preference learning using topic models by implementing the method proposed by Ó Séaghdha and testing it on a monolingual corpus to establish a baseline for evaluating the possible performance gains of using cross lingual knowledge transfer. Evaluation will be done by computing the correlation of the plausibilities output by the model to human plausibility judgements on our previously compiled multilingual set of predicate-argument instances.

Then, the project will build on the work of Peirsman and Pado by attempting to reproduce their results on pairwise, bilingual vector spaces using unrelated corpora: the British National Corpus for English, the TiGer corpus for German and the AnCora corpus for Spanish, followed by investigating the possibility of extending their vector space approach to more than two languages at once. We generalize the bilingual knowledge transfer model to a set of languages Langs using a parameterized Equation 1, where v is a verb, a is an argument, and $P_{L_k}(a|v)$ stands for the plausibility of a as an argument of verb v in a language L_k.

$$P_{L_n}(a|v) = \Sigma_{i \in Langs} \alpha_i P_{L_i}(L_i(a)|L_i(v)) \tag{1}$$

The following step will be to look into possibilities not explored by Peirsman and Pado by testing alternative translation functions and different strategies for choosing the linear parameters α_i. We will test the benefits of using Wikipedia for initializing the bootstrapping stage of the vector space construction process and for evaluating the alignment output, and we will verify the performance of cross-lingual smoothing of plausibilities to quantify the benefits of cross-lingual knowledge transfer.

Within the unilingual models, we will investigate alternatives to the authors' smoothing technique, expressed in Equation 2, by testing other lexical similarity functions, such as those described in the work of Navigli and Ponzetto (2010).

$$P_{L_n}(a|v) = \Sigma_{h' \in Seen(v)} \frac{w(a') sim(h, h')}{\Sigma_{h'} w(h')} \tag{2}$$

Finally, if time allows we will compare our results to those of joint multilingual topic modelling based on the output of the vector space translation, as described in the previous work of Boyd-Graber and Blei (2010) on polylingual topic models. We will use the dimensions of the multilingual vector space as the matching function across vocabularies and we will then try to learn an expanded matching set, the topics and the topic distributions.

The deliverable of this project will be a model for automatically inducing selectional preferences in a multilingual setting along with a system implementing it. Improved performance in comparison to the unilingual models might result in a publication.

3 Detailed Workplan

The project roadmap was divided in fourteen chunks of at most two weeks each, organized as follows:

Chunk 1 (4/12/2012 - 17/12/2012). Compile the batch of evaluation data for all languages used in the project. Create an automated testing framework.

Chunk 2 (18/12/2012 - 31/12/2013). Ensure that all the corpus data is uniformly annotated and preprocessed. Write the literature survey section of the dissertation.

Chunk 3 (3/01/2013 - 15/01/2013). Implement the approach proposed by Ó Séaghdha (2010), which uses monolingual topic models for selectional preference learning. Test the results against the previously compiled data to get a baseline.

Chunk 4 (16/01/2013 - 28/01/2013). Reproduce the results of Peirsman and Pado's bilingual vector space model. Run the implementation with all pairwise language combinations, and quantify the amount of knowledge transfer.

Chunk 5 (29/01/2013 - 11/02/2013). Investigate the extensibility of Peirsman and Pado's vector space model to more than two languages. Test the cross-language plausibility approximation equation for different parameter choices.

Chunk 6 (12/02/2013 - 25/02/2013). Investigate the improvements to building a better-aligned vector space model by using Wikipedia cross-language links to start the bootstrapping procedure.

Chunk 7 (26/02/2013 - 11/03/2013). Investigate alternative plausibility smoothing techniques based on similarity measures acquired from Wikipedia.

Chunk 8 (12/03/2013 - 25/03/2013). Investigate the benefits of cross-lingual smoothing of plausibilities.

Chunk 9 (26/03/2013 - 8/04/2013). Investigate the learning of a cross-lingual mapping function for predicate argument positions, in the presence of small amounts of parsed data.

Chunk 10 (9/04/2013 - 22/04/2013). Investigate the use of joint, multilingual topic modelling using the vector space for supplying translations.

Chunk 11 (23/04/2013 - 5/05/2013). Error analysis.

Chunk 12 (6/05/2013 - 15/05/2013). Thesis writing, part I.

Chunk 13 (16/05/2013 - 01/06/2013). Thesis writing, part II.

Chunk 14 (02/06/2013 - 13/06/2013). Contingencies.

References

[1] Jordan Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 75–82, Arlington, Virginia, United States, 2009. AUAI Press.

[2] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 216–225, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[3] Yves Peirsman and Sebastian Padó. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 921–929, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[4] Diarmuid Ó. Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 435–444, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.