Tagsets Mapping and Statistical Training Data Cleaning-up

Felix Pirvan, Dan Tufis
Research Institute for Artificial Intelligence
Romanian Academy
13, Calea 13 Septembrie, 050711, Bucharest
{felix, tufis}@racai.ro

Introduction

The paper describes a general method (as well as its implementation and evaluation) for deriving mapping models for different tagsets available in existing training corpora (gold standards) for a specific language. These mapping models are further used to significantly improve the accuracy in the underlying training corpora and also for the assessment of the distributional adequacy of various tagsets used in processing the language in case. Unlike other methods, such as those reported in (Brand, 1995) or (Tufis&Dragomirescu, 2004), which assume a subsumption relation between the considered tagsets, and as such they aim at minimizing the tagsets by eliminating the feature-value redundancy, this method is applicable for completely unrelated tagsets. Also, in contrast to previous methods, where the mapping is deterministic, here it is probabilistic. Although the experiments were restricted to morpho-syntactic (POS) tagging, the method is applicable to other types of tagging as well.

Motivations

Let C be a raw text and TAG(C, T) the tagged version of it with tags belonging to the tagset T.

Let us further assume there exist two gold standards (GS): GS(C1) and GS(C2), where Ci is the corpus tagged with the Ti tagset. By using any trainable tagger, one can easily build two language models: LM(GS1) and LM(GS2) based on which a new text Q can be tagged twice: TAG(Q, LM(GS1)) and TAG(Q, LM(GS2)). One could ask various legitimate questions, out of which we were challenged by the following ones:
• which version of Q is more accurately tagged?
• if somebody, for various objective reasons, would like the texts being tagged with one tagset in spite of results systematically showing that the other tagset offers better results, could an automatic mapping ensure better results than tagging with the desired tagset (the less accurate LM)?
• is possible to merge two or more training data using different tagsets and get much more training data for a better language model?
• can annotation errors be detected and automatically corrected in the training corpora?

In processing parallel corpora it is desirable to use for all the languages the same encoding system (e.g. Multext-East, http://nl.ijs.si/ME/, which is underlying all our multilingual resources). However, more often than not, the freely available training data are using different tagsets; having all the data in a similar format would be a very useful thing.

Experimental Data, the Methodology and Evaluation

By direct-tagging of the two gold standards GS(X) and GS(Y) we mean the tagging of one corpus with the LM learnt from the other one: TAG(X, LM(Y)) and TAG(Y, LM(X)). The pairs <TAG(X,LM(Y)) , GS(X)> and <TAG(Y,LM(X)) , GS(X)> are used to derive
the bidirectional tagset mapping model MM(T_X, T_Y), where each tag in T_X is associated with one or more tags from T_Y and vice-versa.

We developed an HMM-based algorithm (details will be provided in the extended paper), called crossLM-tagger, which is more informed than a typical HMM-tagger. It uses two language models, a tagset mapping model, and the direct-tagging of two texts, producing a more accurate tagging of the two texts.

For the evaluation we used the English part of the “1984” multilingual corpus, containing approximately 120,000 tokens (including punctuation), and an excerpt from the Brown corpus of almost the same length. The tagsets used in the two corpora are quite different and in the following we will denote them as T_{1984} and T_{Brown}. The experiments were run several times each time using another fragment of the Brown corpus. We used Brant’s TnT tagger and built the basic language models LM(1984), LM(Brown). By tagging the training corpus with the model extracted from it we got what we call the biased tagging (BT). We conjecture that for a clean training corpus, reasonably large (say, more than 100,000 tokens) and an adequate tagset the accuracy of the tagging should be very high (higher than 98-98.5%). As the accuracy for the BT of the Brown corpus was only 93.8% we expected to find several annotation errors in this training data (for the “1984” corpus the accuracy of BT was 98.7%).

We performed two experiments:

a) the GS(1984) and GS(Brown) were crossLM-tagged and the accuracy gains over the direct-taggings TAG(1984, LM(Brown)) and TAG(Brown, LM(1984)), were 2.05% and 1.68% respectively.

b) the corpora resulted from the first experiment were themselves cross-tagged and the newly resulted corpora were confronted with the gold standards. The 1984 corpus was reproduced in proportion of 98.9% and the Brown corpus excerpt was reproduced in proportion of 96.4%.

By inspection of the systematic differences between the Brown corpus annotation after the second experiment, and its biased tagging, a set of correction rules were extracted. Applying these rules to the double crossLM-tagged Brown corpus, a new improved version of it resulted. The accuracy of this version of the Brown corpus is currently 97.6%. By further analysis and more correction rules added, we are confident that this figure could reach the expected value of 98-98.5%. The current and future versions of the Brown corpus annotation will be publicly released.

Conclusions

Our experiments showed that it is possible to automatically build tagset mapping models and to detect and correct annotation errors in the training data (thus improving the language models). We also showed that biased tagging could give a strong indication of the adequacy of a language model built with a given tagset and our approach can ensure the compromise of having data tagged with the preferred tagset but using a language model based on a more appropriate tagset.

References

Brants, Th. "Tagset Reduction without Information Loss", ACL’95 Conference, Cambridge, Massachusetts, USA, 1995, pp. 287-289

Brants, Th. “TnT- A Statistical Part-of-Speech Tagger” Installation and User Guide, University of Saarland, Computational Linguistics, March 1998.

Tufiș D., Dragomirescu, L. “Tiered Tagging Revisited”. In Proceedings of the 4th LREC Conference, Lisabona, 2004, pp. 39-42