Morpho-syntactic labelling of an oral corpus by decomposing labels

Isabelle Tellier (1), Iris Eshkol (2), Samer Taalab (3), Sylvie Billot (1)

(1) LIFO, université d'Orléans (2) LLL, université d'Orléans

1 Introduction

The Eslo ("Enquête sociolinguistique d'Orléans", i.e. Sociolinguistic Inquiery of Orléans) project allowed to collect a large corpus from transcribed oral interviews. Our purpose here is to build a morpho-syntactic tagger with linear CRF for this corpus, full of disfluences.

2 The corpus and its labelling

A morpho-syntactic tagger associates to each word of a corpus a label which recapitulates its morpho-syntatic properties in the text. In corpora from oral data, not only do we have to face the usual problem of multi-labels words, but also the more specific problems of *disfluences* (repetitions, ungrammatical constructions...), of non existing words and of the lack of punctuation marks [1]. First, inspired by the Cordial tagger, we have defined a new set of morpho-syntactic labels well adapted to oral corpora. These labels are hierarchically defined according to three different levels: a POS level (L0, with 16 different labels), a morphological level (L1, with 72 labels) and a syntactico-semantic level (L3, with 107 labels). Then, we have built a reference corpus for our new set of labels. It has been obtained by using Cordial, then modifying its labels by scripts and manual corrections. The learning corpus finally contains 18 500 words belonging to 1 750 distinct "sequences".

3 The experiments

We use the CRF++ software, where the features are built with the words and their last n ($1 \leq n \leq 3$) characters (noted Dn(word)), supposed to relevant for French morphology. We systematically perform 10-fold cross-validations.

3.1 Reference experiment

The reference experiment consists in trying to directly learn the level L2.

Test	features	Nb features	Accuracy
reference	word, D1(word), D2(word), D3(word)	12 000 000	88%

When we integrate into the features the lemmas provided by Cordial, it increases the accuracy of 2 points, but also increases the learning time.

3.2 Cascade learning

Our first idea to improve the reference experiment is to take advantage of the hierarchical structure of labels, as already done in [2]. To do so, we first learn L0 alone, then integrate the result to learn L1, and finally we integrate the results for L0 and L1 to learn L2. The final accuracy obtained is not better than the previous one (87%), and the learning time is increased. Cascade learning thus does not seem to be worth doing.

3.3 Learning by decomposing and recomposing labels

The labels we have defined can be decomposed into mutually exclusive smaller sets of sub-labels, such that the initial set of labels is included into their cartesian product. For example, the set of possible labels for nouns (N) can be decomposed into: $\{NMS, NMP, NFS, NFP\} = \{N\}.\{M, F\}.\{S, P\}$

We generalize this kind of decomposition to the whole set of labels at level L2: we empirically build four distinct subsets called G0, G1, G2 and G3, and we learn each of them independently:

Test	features	G0=L0	G1	G2	G3
decomposition	word, D1(word),D2(word),D3(word)	92%	92%	95%	94%

But the cartesian product G0.G1.G2.G3 overgenerates: some not relevant labels become possible, for example adding morphological properties to invariable adverbs. To identify a correct label from the components Gi, two distinct approaches are tested. The first one consists in generating a new CRF with the outputs of the four components. The second one consists in taking into account a small set of manually written symbolic rules: for example, if G0 gives ADV as a result, thus the results of the other three components are ignored.

CRF recomposition reaches 87% of accuracy, while symbolic rules recomposition reaches 89%. This is the best result we could obtained: it is the same as the one obtained by taking lemmas into account. But the main gain in this case concerns the learning time, which is reduced to about 75mn instead of 15hours.

4 Conclusion

We managed to learn, with not so many examples, a morpho-syntactic tagger behaving well on a difficul oral corpus. We have shown that CRFs allow to easily integrate linguistic knowledge into their features, and that parallel independent learnings is more efficient than cascading ones.

References

1. C. Blanche-Benveniste, C. JeanJean (1987) Le français parlé. Transcription et édition, Paris, Didier Erudition.
2. F. Jousse (2007) Transformation d'arbres XML avec des modèles probabilistes pour l'annotation, PhD thesis, université of Lille.