

Meta-Learning for Phonemic Annotation of Corpora ¹

Véronique Hoste Walter Daelemans Erik Tjong Kim Sang
Steven Gillis

CNTS Language Technology Group, University of Antwerp

The development of accurate and understandable annotation tools is of prime importance in current Natural Language Processing research, which is based to a large extent on the development of reliable corpora. We discuss the task of phonemic annotation in such a large-scale Dutch-Flemish corpus development project, “Corpus Gesproken Nederlands” (Corpus Spoken Dutch, CGN) in which a 10 million word spoken corpus is collected and linguistically annotated. One of the annotation layers is a representation of the pronunciation of the recorded speech (the phonemic representation). As available speech recognition technology is not yet up to generating this annotation automatically, it has to be produced from the (manually transcribed) orthographic transcription. These pronunciation representations should reflect either Flemish (the variant of Dutch spoken in the North of Belgium) or Dutch pronunciation, depending on the origin of different parts of the corpus. To generate the phonemic representations, we need accurate grapheme-to-phoneme conversion (the module in speech synthesis which converts spelling into phonemic representations). In this study, we investigate the level of accuracy that can be obtained using various machine learning techniques trained on two available lexical databases containing examples of the pronunciation of Flemish (Fonilex, version 1.0b) and Dutch (Celex, release 2). We compare several possible approaches to achieve the text-to-pronunciation mapping task: memory-based learning, rule induction, maximum entropy modeling, combination of classifiers in stacked learning, and stacking of meta-learners.

Training **single classifiers** on both variants of Dutch is performed by making use of TiMBL ², a software package implementing several memory-based learning techniques. The algorithm used for this experiment is called IB1-IG. It extends the basic k-nn algorithm with information gain ratio feature weighting. This already results in generalisation accuracies of 93% at the word level for Celex and 86% for Fonilex which has a more complex phonemic representation (Table 1).

Since we did not have additional information to raise the performance ceiling which can be observed when using data driven systems, we explore whether adding a classifier in the combination having learned a particular task (e.g. Flemish) can help boost performance on a different but similar task (Dutch). Given that the classifiers for Flemish and Dutch are trying to learn very similar but nevertheless

²The TiMBL package can be obtained from <http://ilk.kub.nl/software.html>

Table 1: Results

Classifier	CELEX		FONILEX	
	Word	Phoneme	Word	Phoneme
Single (IB1-IG)	93.00	99.16 (± 0.03)	86.37	98.18 (± 0.04)
Cascade (IB1-IG)	92.90	99.10 (± 0.04)	87.58	98.29 (± 0.03)
Comb.I (IB1-IG)	94.18	99.28 (± 0.03)	88.03	98.36 (± 0.03)
Comb.II (IB1-IG)	95.16	99.40 (± 0.02)	91.55	98.89 (± 0.04)
(C5.0)	93.03	99.16 (± 0.03)	88.41	98.48 (± 0.05)
(IB1-IG)	95.16	99.40 (± 0.02)	91.55	98.89 (± 0.04)
(IGTREE)	94.94	99.37 (± 0.03)	91.33	98.85 (± 0.04)
(MACCENT)	92.07	99.03 (± 0.05)	87.27	98.28 (± 0.04)
(IB1-IG-meta)	95.53	99.45 (± 0.02)	92.25	98.99 (± 0.03)

slightly different mappings, we investigate whether the predicted output of the one could help in making more accurate the predicted output of the other to further improve the accuracy of the grapheme-to-phoneme convertors. The experiments in which single classifiers are trained on Celex and Fonilex, respectively, are taken as the basis for a **cascade experiment** and two **classifier combination** experiments (I and II in Table 1) using the IB1-IG algorithm. In (I) the predicted output for the other variant is used as an additional information source. In (II), the spelling and predicted output for both problems are combined. Table 1 shows that an already high accuracy level for single classifiers is boosted significantly with additional error reductions of 31% for Dutch and 38% for Flemish using combination of classifiers (II).

Going one step further, we **combine combination classifiers**, using four different meta-classifiers, all trained on spelling information together with the output of both single object classifiers, viz. C5.O, which is a commercial version of the C4.5 rule induction program, IB1-IG, IGTREE, which is an optimised approximation of the instance-based learning algorithm IB1-IG and MACCENT, an implementation of maximum entropy modeling.³ The results of these four combination classifiers are used to train a “meta-meta-classifier”, for the training of which IB1-IG is used. The reasoning behind this experiment is that the same way a meta-classifier can overcome some of the errors of different “object-classifiers” learning a similar task, a “meta-meta-classifier” should be able to do the same for meta-classifiers. The use of a “meta-meta-classifier” indeed leads to an additional decrease of 5% for both Flemish and Dutch.

We have shown that the already high accuracy level for single classifiers is boosted significantly when using combination of classifiers and combination of meta-learners. With this high level of accuracy, automatic phonemic conversion becomes an increasingly more useful annotation tool.

³Details on how to obtain Maccent can be found on: <http://www.cs.kuleuven.ac.be/~ldh/>