Statistical Modelling of Highly Inflective Languages

Mirjam Sepesy Maučec
University of Maribor, Slovenia

Zdravko Kačič
University of Maribor, Slovenia

INTRODUCTION

A language model is a description of language. Although grammar has been the prevalent tool in modelling language for a long time, interest has recently shifted towards statistical modelling. This chapter refers to speech recognition experiments, although statistical language models are applicable over a wide-range of applications: machine translation, information retrieval, etc.

Statistical modelling attempts to estimate the frequency of word sequences. If a sequence of words is $s = w_1 w_2 ... w_k$, the probability can be expressed as:

$$P(s) = P(w_1 w_2 ... w_k) =$$

$$\prod_{i=1}^{k} P(w_i | w_1 ... w_{i-1}) \approx \prod_{i=1}^{k} P(w_i | w_{i-n+1} ... w_{i-1}).$$

It is reasonable to simplify this computation by approximating the word sequence generation as a *(n-1)*- order Markov process (Jelinek, 1998). Bigram *(n=2)* and trigram *(n=3)* models are common choices. Although we have limited the context, such models have a vast number of probabilities that need to be estimated. The text available for building the model is called the ´training corpus´ and, typically contains many millions of words. Unfortunately, even in a very large training corpus, many of the possible *n*-grams are never encountered. This problem is addressed by smoothing techniques (Chen & Goodman, 1996).

Which is the best modelling unit? Words are a common choice, but units smaller (or larger) than words can also be used. Word-based *n*-gram is best suited to modelling the English language (Jelinek, 1998). Inflective languages have several characteristics, which harm the prediction powers of standard models.

In general, all Indo-European languages are inflective but a serious problem arises regarding languages which are inflected to a greater extent (e.g. Russian, Czech, Slovenian). Agglutinative languages (e.g. Hungarian, Finnish, Estonian) have even more complex inflectional grammar where, besides inflections, compound words are a big problem. Inflective languages add inflectional morphemes to words. Inflectional morphemes indicate the grammatical information of a word (for example case, number, person, etc.). Inflectional morphemes are commonly added by affixing, which includes prefixing (adding a morpheme before the base), suffixing (adding it after the base), and much less common, infixing (adding it inside the base). A high degree of affixation contributes to the explosion of different word forms, making it difficult, even impossible, to robustly estimate language model probabilities. Rich morphology leads to high OOV (Out-Of-Vocabulary) rates and, therefore, data sparsity is the main problem.

This chapter focuses on modelling unit choice for inflective languages with the aim of reducing data sparsity. Linguistic and data-driven approaches were analyzed for this purpose.

BACKGROUND

Class-Based Language Models

Some words are similar in their morphological, syntactic or semantic functions. In class-based language models, similar words are grouped into classes in order to improve the robustness of parameter estimation:

$$P(w_i | w_{i-1}) = P(w_i | C(w_i)) \cdot P(C(w_i) | C(w_{i-1})).$$

C denotes the deterministic mapping of words into classes. Non-deterministic mapping can also be derived at, where one word can belong to many classes. A model is also applicable, where the word is directly conditioned by the classes of previous words. The idea behind class-based models is parameter-set reduction. There are far fewer free parameters to estimate in a class-based model than in a word-based model.

Words in the same class are similar in a certain way. This similarity can be defined, based on certain external knowledge or statistical criterion. The best known example of clustering using linguistic knowledge is clustering by POS (Part Of Speech). Eight POSs are defined in traditional English grammar: noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. This set of classes is, however, too small for modelling inflective languages. Those classes that reflect additional grammatical features (gender, case, number, tense, etc.) are more suitable.

Linguistic classes were examined for several languages, which are more or less inflective. A language model for French combined POS classes with a component based on lemmas (El-Beze & Derouault, 1990). In the language model for Czech, words were clustered into 410 morpho-syntactic classes (Nouza & Nouza, 2004). 1300 classes were used in another experiment for Czech (Kolar, Svec & Psutka, 2004). Class-based models with linguistic classes also proved to be successful for Spanish (Casillas, Varona & Torres, 2004).

Data driven classes are automatically derived at by statistical means. IBM pioneered this approach (Brown, de Souza, Mercer, Della Pietra & Lai, 1992). In their approach, words are clustered using a greedy algorithm that tries to minimize the loss of mutual information between classes incurred during the merge. The number of classes must be defined in advance. The algorithm continues to merge pairs of classes until the desired number of classes has been obtained. Another greedy approach uses the exchange algorithm (Martin, Liermann & Ney, 1995). Each word is moved from its class to another one if it maximizes mutual information between classes.

Data-driven class-based language models have been built for many inflective languages. For French they show improved performance on small and large corpora (Zitouni, 2002). The results have been improved by using a hierarchical language model with variable-length class sequences, based on 233 grammatical classes. In experiments on the Russian language, the best results were obtained by using 500 classes (Whittaker & Woodland, 2003). The results were further improved when a class-based model was combined with a word-based model.

Lots of data must be available to derive at classes automatically from the data instead of using external knowledge sources.

Language Models Based on Sub-Word Units

Given the difficulties in language modelling based on full word forms it would be desirable to find a method of decomposing word forms into their morphological components and to build a more robust language model based on probabilities involving individual morphological components.

Lexicons exist for some languages which contain information about the morphological components of words. In experiments on Czech, words were decomposed into stems and endings using a Czech Morphological Analyzer, and were then used as modelling units (Byrne, Hajič, Ircing, Krbec & Psutka, 2000). Morpheme-based language models were also studied for the Korean language, where a word-phrase is an agglomerate of morphemes (Kwon & Park, 2003). Sub-word units are also used when modelling agglutinative languages where, besides inflections, compound words are very common (Szarvas & Furui, 2003). Morphological sub-word units have also been proved for Turkish (Erdoğan, Büyük & Oflazer, 2005). The language model's constraints were represented by a weighted finite state machine.

Many languages do not have developed morphological analysers. Data-driven discovery of a language's morphology is used in such cases. It is common for data-driven approaches to outperform linguistic ones. Morphemic suffixes were discovered by Minimum Description Length (MDL) analysis (Brent, Murthy & Lundberg, 1995). MDL analysis has been used for morphological segmentation for various European languages (Goldsmith, 2001). An algorithm for learning morphology using latent semantic analysis was also discovered (Schone & Jurafski, 2000). This algorithm only extracts affixes when the stem and stem-affix are sufficiently similar semantically. The language model for Russian also improves when using data-driven sub-word units (Whittaker & Woodland, 2000). Language-independent algorithms for discovering word