The reliability of statistics in linguistics
Notes to a dictionary extension

Mátyás Naszódi

MorphoLogic
1122 Budapest, Ráth György utca 36.
naszodim@morphologic.hu

Abstract. Nowadays statistical tools are often used tool in linguistics, but the reliability of these methods is rarely examined. In natural language processing, statistical methods have their boundaries, and one should pay more attention to them. I try to show, when and how can we estimate its boundaries.

1 Uncertainty in languages

Due to the natural features of the languages, there are many types of uncertainty. To decide whether a word form is correct or not is sometimes questionable. The syntax of a language depends on its creator, and even linguists are unconvinced about the correctness of certain sentences. It is impossible to find the only right translation; there are, however, bad, good and better translations of a text. Probabilistic models can describe the problems in all cases.

In spite of the uncertainty, a user would hate a spell checker that marks words with “perhaps”, “sure”, or percentage of correctness; or would hate a translation tool offering hundreds of possible solutions.

2 Characteristics of linguistic statistics – Zipf Law

If a linguistic phenomenon has more then thousand distinguishable classes, the distribution of the phenomena by the row of classes show similar characters. It is described by the Zipf Law. It tells that the probability of a class and the order in distribution row are in correlation:

$$f(n) \approx C/n^s \quad (1)$$

where n is the ordinal in the probability order of the n^{th} class, C is a constant for normalizing the equation, s is an exponent that is a little bit less than 1 [1].

For low n, that is for classes of high probability, the estimation is far from correct. For classes of low probability, the estimations are biased by measuring errors. In the

middle part of the row however, the Zipf Law works well. In linguistic cases the exponent s is as nearer to 1 as larger the number of classes.

3 Mathematics of quality in linguistic works

While collecting or processing a linguistic database (creating a dictionary), the time (of the work) for the collection of N items might be in linear correspondence with the number of items, if the work is homogenous in the set of items.

$$T(N) = \sum_i e_i = N*e \tag{2}$$

If you take it in account that rare items need larger corpora to find them in it, and they need more time to code them, the equation should be changed:

$$T(N) = \sum_i e_i = C*\sum_i i^s = C* N^{1+s}/(1+s) > O(n^2) \tag{3}$$

The quality of coding of individual items gets worse by the rarity of the items. For gaining a quality level, the necessary time is fast growing with the requirement of quality:

$$T(q) > O(1/q^2) \tag{4}$$

where $1-q$ is the covering, that can be a measure of quality.

If the time for preparing an item is limited, the quality of the work gets worse by the number of items. In that case, the (dictionary) work loses required quality in a well-defined number of items. That is why a dictionary may reach its optimal size, and machine translation based on memory or statistics reaches quickly its maximal quality level.

Quality barrier may be broken by independent evaluations and reduction of the number of classes. (It may decrease the constant s of the Zipf equation). An example for the first one is when parallel coding is used for better quality. Here are some examples for the second one:

1. User dictionaries (spelling checkers, dictionaries)
2. Thematic terminologies (spelling checkers, dictionaries, translators)
3. Morpheme-based statistics instead of wordform-based one for translations

I tried to estimate the value s in translations of the same text to several languages. The results cause some surprise because of the following reasons:

1. The corpora are too small.

2. The measured numbers depend not only on the languages, but on the novel and on the translation as well.

3. Coding errors also biased the data.

Despite of that, data show that in case of languages where the number of word forms is large, the probability of word-forms is nearer to the reciprocal value than in languages with poor morphologies.