Morphological learning as principled argument

Lars G Johnsen
Dept. Linguistics and literature
University of Bergen
lars.johnsen@lili.uib.no

Abstract

We develop a morphological learner that evaluates evidence supporting specific claims that a string of letters is a distributional meaningful unit. The distributional evidence is evaluated by selectional properties of morphs, while evidence towards meaning is modelled by looking at the relationship between stems and words. To assess a proposed affix, it gets a probability measure of meaning by comparing all the possible stems the affix occur with to the particular subset that also occur as words. Since for a stem to be a word counts as evidence towards its meaning, the ratio formed by taking stems that are words to the whole set of possible stems for an affix gives a predictive probability measure for the affix that measures the chance that it has combined with a meaningful stem. This measure, taken in conjunction with the selectional statistics of stems and affixes, provides a basis for deciding on the best morphological structure for a given word. The results for English show a combined precision and recall of 45.

1 Introduction

A lexicon for a language will contain, among a lot of facts about the language, a list of words, the morphemes and rules for how to combine different morphemes like stems and affixes into words. It is assumed that the rules are those of concatenative morphology. Given a lexicon with the above properties, the following statement[1] captures the conditions on a word *w* that consists of a stem *x* plus suffix *y*.

$$morph(w, x, y) \Leftrightarrow stem(x) \wedge suff(y) \wedge sel(x, y) \quad (1.1)$$

The binary predicate *sel* encodes the selectional restriction between *x* and *y*. Joining two morphs together may not result in a well formed word, and *sel* encodes the pairs that can be combined together.

Depending on the language, there will be a couple of rules like those in equation (1.1): one for prefix plus stem, stem plus suffix, stem plus infix, and stem plus stem. For the purpose of the morphochallenge task, here restricted to English, we do not consider infixes, nor reduplicative morphology or suprasegmental morphology.

The definition in equation (1.1) splits a word only one time. In order to get a list of morphs from this definition it has to be applied recursively. A predicate *morphs* serves this purpose, and can be defined as follows[2], relating a word form *w* to the list of morphs constituting *w*, in this case a '+' separated list:

$$\begin{aligned} &morphs(w, morphlist) \\ &\Leftrightarrow \\ &morph(w, stm, suff) \\ &\wedge \\ &morphs(stm, stemlist) \\ &\wedge \\ &morphlist = stemlist + suff \end{aligned} \quad (1.2)$$

Our point of departure is that the characterization of morphological analysis is the same

[1] Standard notation from predicate logic is used.

[2] The defining expression translates into the Prolog programming language.

whether the lexicon is given or not. The difference between a learner of a lexicon and a knowledgeable performer is viewed as a difference in the level of confidence.

The problem of learning a lexicon from a word list is, according to this view, taken to be the problem of estimating the truth of the terms in equation (1.1). The truth is assessed via probability measures over a set of hypotheses. A simplifying assumption is that each word has a unique split into stem and affix.

Following (Goldsmith, 2001) the number of hypothesized suffixes considered for an English word *w* is limited to six including null morphs. The set of suffix hypotheses for a word like *drinking* is then

$$\begin{aligned} H = \{ & \\ & morph(drinking, drinking, \emptyset), \\ & morph(drinking, drinkin, g) \\ & morph(drinking, drinki, ng) \\ & morph(drinking, drink, ing) \\ & morph(drinking, drin, king) \\ & morph(drinking, dri, nking) \\ \} & \end{aligned} \tag{1.3}$$

The present approach explores ways for calculating the morphological structure using only the distributional properties of stems and suffixes considered as atoms. Their internal letter structure is not taken into account, but see e.g. (Goldsmith, 2001; Creutz & Lagus, 2005) for how one may go about using that kind of information. The information contained in the inherent substring ordering of the morphs is not utilized either.

2 The probability formulation

Equation (1.1) contains the logical statement of the relationship between a stem and an affix conditioned on the facts in the lexicon. This is converted into a probability equation conditioned on the wordlist *W* considered as a set of propositions of what counts as a word.

$$\begin{aligned} & p(morph(w, x, y) \mid W) = \\ & p(stem(x), suff(y), sel(x, y) \mid W) \end{aligned} \tag{1.4}$$

The right hand side can be expanded to a product of the terms in (1.5) and (1.6) below.

$$p(stem(x), suff(y) \mid sel(x, y), W) \tag{1.5}$$

and

$$p(sel(x, y) \mid W) \tag{1.6}$$

When replacing the lexicon with the wordlist *W* as the conditioning facts in these equations, a couple of assumptions have to be revised. The use of *sel* as a conditioning term in equation (1.5) is assumed to be superfluous. The predicate *sel* is for a learner reinterpreted as continuous measure of selectional information, and as such really is a ternary predicate, relating two possible morphs to their selectional information. By doing this, *sel* contributes to the overall value solely through equation (1.6).

This independence assumption turns (1.5) into (1.7) below

$$p(stem(x), suff(y) \mid W) \tag{1.7}$$

The probability formulation then leaves us to compute the equations (1.6) and (1.7).

We will make one change to the objects in the equations. Instead of working with the morph tokens themselves, they are replaced with their respective distributions, indicated using a * on the morph variable.

A stem *x* corresponds then to the class of possible suffixes it combines with. In the following equations a dot "." is used to indicate concatenation.

$$x \mapsto x^* = \{ z \mid x.z \in W \} \tag{1.8}$$

A suffix *y* corresponds to the class of stems it combines with

$$y \mapsto y^* = \{ z \mid z.y \in W \} \tag{1.9}$$

2.1 Selection

The selectional properties are computed by comparing *W* with the possible combinations from *w=x.y* of stems from *y** and suffixes from *x**, denoted y*.x*. This object is closely related to the paradigm in (Snover & Brent, 2002) and the signatures in (Goldsmith, 2001).

The conditional probability of the two sets *y*.x** and *W* is interpreted in a standard way as being the proportion of successes of their intersection, which is computed as the ratio of good words from *y*.x** to all words in *y*.x**.

$$p(sel(x, y) \mid W) = \frac{| y^*.x^* \cap W |}{| y^*.x^* |} \tag{1.10}$$

The implementation used in the morphochallenge uses a *beta*(*a,b*) density for calculating this equation. The first argument, *a*, of this distribution is filled with the positive cases, the numera-

tor of (1.10), and the second argument, *b*, consists of the number of negative cases, the difference between denominator and numerator. The probability assigned to the selectional property is calculated from this density by taking its mean and subtracting one standard deviation. Subtracting one standard deviation gives a more conservative predictive probability than the taken from (1.10) directly, and will penalize those combinations that contain few examples.

For some stems and affixes, the total number of possible words got rather large, and so for the challenge, there was some experimentation with reducing the parameters for the beta density. Best results for both precision and recall was achieved by shrinking the parameters *a* and *b* by the 6^{th} root.

Future improvements for the computation of *sel,* rest on a Bayesian inversion of the formula (1.10), which can be used in updating a distribution *d* over stems and suffixes via maximum likelihood. A particular *d* is a distribution over the hypotheses for a word as shown in (1.3).

The following equation lets *d* play a role in the computation of *sel* as well, and allows us to take into account various confidence levels as expressed by *d* in particular analyses.

$$p(d \mid sel(x, y), W) = \\ = \frac{p(sel(x, y) \mid W, d) \cdot p(d \mid W)}{p(sel(x, y) \mid W)} \quad (1.11)$$

The denominator and normalizing constant of the right hand side of this equation corresponds to the left side of (1.10) and can by using (1.11) be computed by summing over the relevant distributions *d*

$$p(sel(x, y) \mid W) = \\ = \sum_d p(sel(x, y) \mid W, d)\, p(d \mid W)$$

2.2 Stem and affix

There is no source of meaning for stems and affixes from the word list *W* beyond the assumption that any word itself has a meaning. We exploit this fact in the evaluation of the term (1.7) which we will rewrite slightly. Instead of expanding (1.7) into one term for computing the stem and one for the affix, we make the assumption that any evidence that the possible stem is a stem, also counts as evidence that the putative affix is an affix, and vice versa. Accordingly, the two propositions are combined into one, so that (1.7) becomes

$$p(stemsuff(x, y) \mid W) \quad (1.12)$$

This term gets its value solely from the assessment of meaning as follows. The stem *x* is in the word context *xy* so we ask what the probability is that *x* has any meaning given this contextual information. This is turned into an issue of predictive probability: what is the chance of finding a meaningful string in front of *y*? For example, in English, what is the probability for a token to have meaning in the context

x.ing=[open.ing, str.ing, s.ing, r.ing, laugh.ing, talk.ing]?

Three of the *x's* have an independent distribution on their own, namely [open, laugh, talk], so out of these six, the chance is 50% for anything picked out in front of *ing* is a standalone word and carrying meaning using this measure. Note that the stem itself in stem affix combination is not evaluated directly. The independent distribution of stems is used in classifying the affix which in turn classifies the stem.

A predictive probability measure formulated on the basis of the foregoing discussion is then the ratio of actual words in *y** to *y**.

$$p(stemsuff(x, y) \mid W) = \\ = \frac{| y^* \cap W |}{| y^* |} \quad (1.13)$$

As for the case of selection, a beta density is used to localize this ratio. The actual probability assigned takes the standard deviation of this density into account in the same way as for the selection. Affixes with low frequency is penalized by this way of calculating the probability.

A crucial assumption for this approach to work is that the empty suffix is a witness for meaning through the word list. A mild supervision can be built into the learner by supplying other witness morphs that can be used as context for a possible stem. Using the word list as a witness set for meaning presupposes that a good portion of stems are actual words, enough so that different affixes can be distinguished on the basis of it.

With access to a corpus a better model of meaning can be formulated, as shown in (Schone & Jurafsky, 2000).

2.3 Combining the results

The probabilities from each of these estimates are combined for each hypothesis resulting in a

total score, and ranking of all the hypotheses. The decision scheme adopted is to select the best hypothesis, i.e. the one with highest probability. An alternative method could be iterative: remove the worst and recalculate the probabilities, and repeat that process until only one hypothesis remains.

Selecting the highest ranked hypothesis results in an F-score of 45%, recall at 54%, and precision at 39% for the English word list, using the tools available for the competition.

3 Conclusion and further work

We have shown how one can use the concept of meaning in evaluating the different candidates for morphological analysis. The method should lend itself to all languages that permit a certain proportion of its stems to occur as words.

The work reported here is in a state of flux and particularly equation (1.11) is explored.

Acknowledgements

I would like to thank Christer Johansson and Kolbjørn Slethei for discussions on the matters presented here, and three anonymous reviewers for their comments on an earlier draft. Part of this work was supported by a grant from the Norwegian research council over the KUNSTI program, project BREDT.

References

Creutz, M. & Lagus, K. (2005). Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (pp. 106-113). Espoo.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics, 27,* 153-198.

Schone, P. & Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *CoNLL-2000 and LLL-2000* Lisbon, Portugal.

Snover, M. & Brent, M. (2002). A Probabilistic Model for Learning Concatenative Morphology. In *Proceedings of NIPS 2002*.