Unsupervised Morphological Disambiguation using Statistical Language Models

Mehmet Ali Yatbaz and Deniz Yuret

Dept. of Computer Engineering

Koç Üniversitesi

Introduction

•The morphological disambiguation can be defined as the selecting the correct parse of a word in a given context from the possible candidate parses of the word.

• The main challenge of the supervised morphological disambiguation is the difficulty of acquiring a sufficient amount of consistent morphologically parsed training data.

•Another issue is, unlike English, in agglutinative languages the number of theoretically possible parses can be infinite although the number of features is finite.

Below you can see three possible morphological parses for the Turkish word "masalı".

Stems	Morphological Parses	Meaning
masal	+Noun+A3sg+Pnon+Acc	(= the story)
masal	+Noun+A3sg+P3sg+Nom	(= his story)
masa	+Noun+A3sg+Pnon+Nom^DG+Adj+With	(= with tables)

Unsupervised Morphological Disambiguator

Model

• The main idea of our model is it assigns parses to the contexts instead of words itself.

• Thus, our model selects the parse t of the target word w that is most likely in the target word context, c_w.

• To achieve this, the model finds t that maximizes $P(t|c_w)$ using the replacement words from the vocabulary, V.

$$\arg\max_{t \in T_w} P(t \mid c_w) = \sum_{v \in V} P(t \mid v, c_w) P(v \mid c_w)$$

Estimation

$P(v|c_w)$ is estimated using the n-gram language model trained on a 400 million words Turkish web corpus.

• c_w is defined as the 2n–1 word window $w_{-n+1}\dots w_o \dots w_{n-1}$.

• Finally,

$$\begin{aligned} P(w_o = v) &\propto P(w_{-n+1} \dots w_0 \dots w_{n-1}) \\ &= P(w_{-n+1}) P(w_{-n+2} \mid w_{-n+1}) \dots P(w_{n-1} \mid w_{-n+1}^{n-2}) \\ &\propto P(w_0 \mid w_{-n+1}^{-1}) \dots P(w_1 \mid w_{-n+2}^{0}) \dots P(w_{n-1} \mid w_0^{n-2}) \end{aligned}$$

$P(t|v,c_w)$ is estimated using two assumptions

1. **Pruning assumption:** Every w has a possible parse set T_w. Parses that are not in T_w have zero probability in the context of w.
2. ***Uniformity assumption:*** The distribution of parses given a replacement word v and context c_w is uniform on T_w.

$$P(t \mid v, c_w) = \begin{cases} \dfrac{1}{|T_w \cap T_v|} & t \in T_w \cap T_v \\ 0 & otherwise \end{cases}$$

Parse Simplification

•The estimation quality of $P(t|c_w)$ highly depends on the parse T_w.

• Instead of using the parses directly we construct a discriminative minimal set S_w by selecting the minimum number of rightmost features of each parses.

Stems	Morphological Parses	Simplified Parses
masal	+Noun+A3sg+Pnon+Acc	Pnon+Acc
masal	+Noun+A3sg+P3sg+Nom	P3sg+Nom
masa	+Noun+A3sg+Pnon+Nom^DG+Adj+With	With

Algorithm

1. Construct a morphological dictionary for all the words in V.
2. Construct S_{w_i} by simplifying T_{w_i} where w_i is the i^{th} target word.
3. Calculate $P(v_{ij}|c_i)$ where v_{ij} is the j^{th} replacement of w_i.
4. Calculate $P(t|c_i)$ for all t in S_{w_i} using the probabilities calculated in Step 3.
5. Select t that maximizes $P(t|c_i)$.

	Test Set	Tagged Trained Set
Sentences	446	50673
Tokens	5365	948404
Ambiguous Tokens	45.4%	42.1%
Average Parses	1.85	1.76

Experimental Results

We define an unsupervised and a supervised baseline.

1. **Unsupervised Baseline:** Randomly pick a parse of w from T_w. Disambiguate **39.4%** of the ambiguous words.
2. **Supervised Baseline:** Select a parse of w from T_w by using majority voting. Disambiguate **71.0%** of the ambiguous words.

Effect of Corpus Size on our model:

We used three corpora with different sizes to train 4-gram language model. We randomly select 1% and 10% of the original training corpus.

Corpus Size	Accuracy
4M	60.4
40M	63.1
400M	64.5

As the corpus size becomes smaller, the accuracy of the model decreases significantly (in terms of 95% confidence interval). Thus, the performance of the model can be improved by using a larger Turkish corpora.

Effect of Replacement Word Number on our model:

We calculate $P(v|c_w)$ of each replacement word and select 10, 100, 200 and 2000 replacement words that have the highest $P(v|c_w)$ and use only these words to estimate $P(t|c_w)$.

Number of replacements	Accuracy
Top 10	63.4
Top 100	64.3
Top 200	64.4
Top 2000	64.5

This experiment shows instead of calculating $P(v|c_w)$ for all vocabulary, top k $P(v|c_w)$ values can be used since the results are not different (in terms of 95% confidence interval).

Conclusion

• Our model assigns parses to context instead of assigning them to words.

•The probabilities of morphological analysis are calculated using a language model. Therefore it can be applied to any language without predefining any language dependent rules.

•**We were able to achieve 64.5% accuracy.** This accuracy might be improved by relaxing the uniformity assumption and letting it to converge to the actual probabilities.