Statistical Structure in Language Processing
Phrase based models

Cristina Gârbacea
10407936
cr1st1na.garbacea@gmail.com

Sara Veldhoen
10545298
sara.veldhoen@student.uva.nl

Abstract

Bla bla bla

1 Introduction

2 Phrase Extraction and weight estimation

In this section, we present our approach to the extraction of phrase pairs from the corpus. Subsequently, we

Phrase Extraction The number of possible phrase pairs per sentence pair is huge: each sentence can be partitioned in a vast amount of ways, and each partition could form a phrase pair with any partition in the paired sentence.

In order to reduce the space, we consider only phrase pairs that are consistent with the alignments produced by IBM models. As in (1), consistency is defined as follows:

$$\begin{aligned} \langle \bar{e}, \bar{f} \rangle \text{ is consistent with } A \Leftrightarrow \\ \forall e_i \in \bar{e} : \langle e_i, f_j \rangle \in A \Rightarrow f_j \in \bar{f}, \\ \text{and } \forall f_j \in \bar{f} : \langle e_i, f_j \rangle \in A \Rightarrow e_i \in \bar{e}, \\ \text{and } \exists e_i \in \bar{e}, f_j \in \bar{f} : \langle e_i, f_j \rangle \in A. \end{aligned}$$

For this assignment, the symmetrized alignments of the corpus sentences were given. We base our extraction algorithm on the one presented in (1, page 133). We iterate over all windows up to a certain length in the English sentence, and find the foreign windows that are consistent given the alignment. For all valid pairs of windows, we extract the corresponding phrase pair.

Conditional Probability Estimates

Joint Probability Estimates In

3 Experiments and Results

4 Conclusion

References

[1] Philipp Koehn, 2010. *Statistical Machine Translation.* Cambridge University Press.

[2] Daniel Marcu and William Wong, 2002. *A phrase-based, joint probability model for statistical machine translation.* Association for Computational Linguistics.