Aligning Turkish and English Parallel Texts for Statistical Machine Translation

İlknur D. El-Kahlout and Kemal Oflazer

Faculty of Engineering and Natural Sciences, Sabancı University,
Istanbul 34956, Turkey
ilknurdurgar@su.sabanciuniv.edu, oflazer@sabanciuniv.edu
http://www.hlst.sabanciuniv.edu

Abstract. This paper presents a preliminary work on aligning Turkish and English parallel texts towards developing a statistical machine translation system for English and Turkish. To avoid the data sparseness problem and to uncover relations between sublexical components of words such as morphemes, we have converted our parallel texts to a morphemic representation and then used standard word alignment algorithms. Results from a mere 3K sentences of parallel English–Turkish texts show that we are able to link Turkish morphemes with English morphemes and function words quite successfully. We have also used the Turkish WordNet which is linked with the English WordNet, as a bootstrapping dictionary to constrain root word alignments.

1 Introduction

Availability of large amounts of so-called parallel texts has motivated the application of statistical techniques to the problem of machine translation starting with the seminal work at IBM in the early 90's [1,2]. Statistical machine translation views the translation process as a noisy-channel signal recovery process in which one tries to recover the input "signal" e, from the observed output signal f.[1] Thus given some output sequence f one tries to find

$$e^* = \arg\max_e P(e|f)$$

as that (English) sentence that maximizes the probability of giving rise to the specific output (French) sentence f. Using Bayes' law, this probability is expanded into

$$e^* = \arg\max_e P(e|f) = \arg\max_e \frac{P(f|e)P(e)}{P(f)} = \arg\max_e P(f|e)P(e)$$

since f is constant for all candidate e's. This formulation has two components: the first component called the *translation model* gives the probability of translating e

[1] Denoting *English* and *French* as used in the original IBM Project which translated from French to English using the parallel text of the Hansards, the Canadian Parliament Proceedings.

P. Yolum et al.(Eds.): ISCIS 2005, LNCS 3733, pp. 616–625, 2005.

into f and the second component called the *language model* assigns the sentence e, a certain probability among all possible sentences in the source language.

Early statistical machine translation systems used a purely word-based approach without taking into account any of the morphological or syntactic properties of the languages [2]. Later approaches exploited morphology and/or syntactic properties in one way or the other, to increase the quality of parameters for the translation model and also to rely on smaller parallel texts [1,3,4,5].

The translation model relies on model parameters that are estimated from sentence-aligned parallel texts [2]. Obviously, for accurate estimation of parameters, one needs large amounts of data which for some language pairs may not be easy to obtain. This can be further complicated by the nature of the languages involved as may be the case for the Turkish and English parallel texts. Even a cursory analysis of sentence aligned Turkish and English texts indicates that translations of certain English words to surface as various morphemes embedded into Turkish words. Thus for accurate estimation of parameters, one needs to consider sublexical structures.

In this paper, we present results from aligning Turkish and English parallel texts towards developing a translation model from English to Turkish for use in a statistical machine translation system. We use morphology in a similar way to Lee [4], but with further exploitation of allomorphy to get more accurate statistics and use a Turkish WordNet [6] that is aligned with the English WordNet [7] as a dictionary for root word alignment.

This paper is organized as follows: we start with a short overview of Turkish morphology to motivate its impact on alignment with English texts for deriving translation model parameters. We then present results from aligning Turkish texts with English texts, followed by the use of the aligned Turkish and English WordNets as a constraining dictionary to improve translation model parameters. We conclude by discussing future work that will make use of this translation model.

2 An Overview of Turkish Morphology

Turkish is an Ural-Altaic language, having agglutinative word structures with productive inflectional and derivational processes. Turkish word forms consist of morphemes concatenated to a root morpheme or to other morphemes, much like "beads on a string". Except for a very few exceptional cases, the surface realizations of the morphemes are conditioned by various regular morphophonemic processes such as vowel harmony, consonant assimilation and elisions. The morphotactics of word forms can be quite complex when multiple derivations are involved. For instance, the derived modifier `sağlamlaştırdığımızdaki`[2] would be broken into surface morphemes as follows:

[2] Literally, "(the thing existing) at the time we caused (something) to become strong". Obviously this is not a word that one would use everyday. Turkish words (excluding noninflecting frequent words such as conjunctions, clitics, etc.) found in typical running text average about 10 letters in length. The average number of bound morphemes in such words is about 2.