Two New Models of Target Language Morphology in Translation

PI: Chris Dyer
Carnegie Mellon University, School of Computer Science
5000 Forbes Ave., Pittsburgh, PA 15213, USA
+1-240-328-4439, cdyer@cs.cmu.edu

Abstract

This proposal addresses the problem of translation into morphologically rich languages with two new models. In the first, we generate new word types that are compositional translations of multiple source words (e.g., compounds in German) and augment existing translation models with these. In the second, we propose using automatically learned distributed representations of morphemes and their contexts instead of hand-engineered feature sets.

1 Project Overview

Complex word formation processes challenge the conventional assumptions made in statistical models of language. When translating into morphologically rich target languages, these assumptions are particularly acute and cause two substantial problems: (i) the translation rule inventories lack the full complement of forms that should be available in the target language and (ii) for the words that are present, model parameters are poorly estimated due to naïve independence assumptions—namely, that different word types do not share statistical strength. We propose work that will address both (i) and (ii) simultaneously by decomposing the translation problem into a two-step process where first a set of candidate lexical/phrasal translations are generated using a morphologically-aware process, then these are used to translate the sentence. In previous work, my research group has demonstrated that this decomposition is effective for dealing with problems as varied as generating non-translating function words, spoken language translation, and dealing with inflectional morphology (Chahuneau et al., 2013b; Tsvetkov et al., 2013, 2014).

This proposal extends our previous work in two ways. First, we consider the problem of producing morphologically complex words that compositionally express the meaning of multiple source words (§1.1). Example phenomena include productive compounding processes in languages like Dutch, Swedish, and German (Declaration of Independence → Unabhängigkeitserklärung), and rich derivational processes such as found in Finnish, Hungarian, and Turkish (I could not understand → anlayamadım). Second, we seek to leverage the fact that morphological paradigms are often multi-dimensional (e.g., a single Russian suffix expresses the gender, number, and tense of the subject) by learning distributed representations of the morphemes we are predicting and the contexts we are conditioning on (§1.2).

This work will be translation from morphologically poor English into several morphologically rich languages. We do so for two reasons. First, as of April, 2014, an estimated 56% of content on the web is published in English. Second, while we wish to make no assumptions about resources in the target language (beyond the availability of monolingual and parallel data), English has mature tools for syntactic analysis that can provide valuable features for predicting target side morphology (Avramidis and Koehn, 2008; Chahuneau et al., 2013b; Yeniterzi and Oflazer, 2010).

In contrast to most previous work, the baseline translation options, features, and decoding algorithms (and systems) are preserved, but morphological processes are modeled in a preprocessing step, making these techniques effective for improving large-data, state-of-the-art systems (Ammar et al., 2013).

1 http://w3techs.com/technologies/overview/content_language/all
1.1 Translating Multiple Source Words into a Single Target Word

Part 1 of the proposed work considers how to deal with the case when multiple source words should be translated into a single target word. Consider the following example:

Input: Washington is considering a weapons delivery to Ukraine.

Google Translate: Washington erwägt eine Waffen Lieferung an die Ukraine.

Desired: Washington erwägt eine Waffenlieferung an die Ukraine.

Presumably, the phrase weapons delivery was not available in the parallel training data. However, it translates completely compositionally to Waffenlieferung, so we should be able to synthesize this as a translation option. In previous work, we inflected single words with different grammatical features (case, number, gender, etc.). To deal with the multi-source case, we propose the following. First, we segment complex words in the target side of the parallel training data (e.g., Waffenlieferung → Waffen Lieferung) using an unsupervised nonparametric Bayesian model (Chahuneau et al., 2013b) and align these to the English translations. Using the multi-source translations that we identify via their alignments, we will train a tree-structured discriminative model (Cohn and Blunsom, 2005) to identify multiword (MW) units that are likely to translate as a group, and train morpheme translation and language models to generate sequences of morphemes input collections of source words and syntactic contexts. At test time, a candidate list of multiword units will be extracted from each source English sentence (together with its parse tree) and then these will be translated using the morpheme translation model, stitched together to form an inflected word, and then these will augment translation options extracted using the usual heuristics.

For part 1, we will evaluate on translation into Turkish, German, and Arabic—three typologically distinct languages in which multi-source word translation plays an important role. Final system quality will be assessed with BLEU, METEOR, and TER and compared to two baselines: one with no awareness of morphological processes and one that translates entire sentences at the morpheme level. Performance of the MW identification model will be assessed with precision-at-N and likelihood on held-out data; performance of the morpheme level translation model will be assessed with morpheme BLEU and likelihood.

1.2 Distributed Representations of Context and Morphology

The second thrust of our proposed work will be to improve how morphemes are modeled. We argue that distributed representations, which represent complex objects as vectors and which are learned automatically from data, are a particularly natural fit for the problem of modeling morphology—in fact, the paradigm tables familiar from textbooks using in foreign language classes are a kind of distributed representation of morphology. In our previous work (Chahuneau et al., 2013b), we developed a model that linked hand-engineered input representations ($x \in \mathbb{R}^n$) to hand-engineered output representations ($y \in \mathbb{R}^m$) using a bilinear form, $xW^\top y$, where $W \in \mathbb{R}^{nxm}$ was learned to maximize likelihood of the training data. Although we were able to leverage a great deal of training data, $W$ was impractically large and the model was difficult to regularize. Since we expect that inflectional features tend to live in a relatively low-dimensional

---

3To keep the project to a reasonable scope, we will only generate words observed in monolingual training data. The monolingual data available is quite large and this decision lets us avoid the open vocabulary language modeling problem, which we have addressed elsewhere (Chahuneau et al., 2013a).

4This component is reminiscent of the often-proposed technique of translating into a morphologically segmented variant of the target and then stitching them together to generate inflected forms (Clifton and Sarkar, 2011; Fraser et al., 2012; Stymne et al., 2013, inter alia); however, we emphasize that this process augments the usual direct translation of inflected forms rather than replacing the entire translation process.
space (e.g., they reflect grammatical notions such as gender, number, or case), we think a more direct approach is to enable the model to learn its own representations along with $W$. In this phase 2, we will focus on translation into Russian in addition to the languages discussed above. Russian is typical of languages with fusional morphology where a single morpheme expresses multiple grammatical features (e.g., -юю “noncompositionally” expresses accusative case, singular number, and feminine gender). In addition to the phase 1 evaluations, we will additionally compare the learned morpheme representations to those available from rule-based morphological analyzers.

---

**References**


Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. Modeling inflection and word-formation in SMT. In *Proc. EACL*, 2012.


