Improving Statistical MT through Morphological Analysis

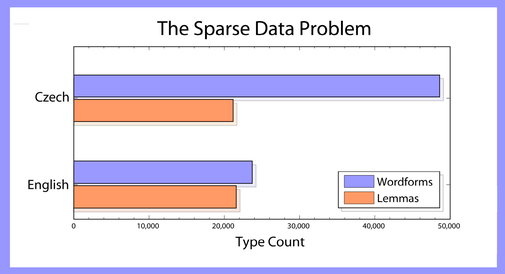
Sharon Goldwater
Cognitive and Linguistic Sciences
sharon_goldwater@brown.edu

David McClosky
Computer Science
dmcc@cs.brown.edu

BROWN

Motivation

In statistical machine translation, data sparsity can make it difficult to estimate word-to-word alignments. In morphologically complex languages, words may be highly inflected. This leads to lower average token frequencies and more sparse data than in less morphologically complex languages. *How much can morphological information alone improve machine translation?*



The sparse data problem in Czech is illustrated by graphing the number of lemma and word types with fewer than 5 occurrences in our corpora. The number of infrequent lemmas is similar in English and Czech, but Czech contains about twice as many infrequent wordforms.

Corpus

We used data from the Prague Czech-English Dependency Treebank.

- Language model trained on ~50k sentences of Wall Street Journal.
- Translation model trained on ~21k parallel sentences of Wall Street Journal.
- Evaluation is on ~250 sentences, each with five human translations.
- Czech words are morphologically annotated with lemmas and up to 15 morph tags for features such as part of speech, tense, gender, etc.

Word	Lemma	Morph tags
Tyto	tento	PDIP1----------
návrhy	návrh	NNIP1-----A----
společnost	společnost	NNFS1-----A----
v	v	RR--6----------
minulosti	minulost	NNFS6-----A----
odmítla	odmítnout	VpQW---XR-AA--1
.	.	Z:-------------

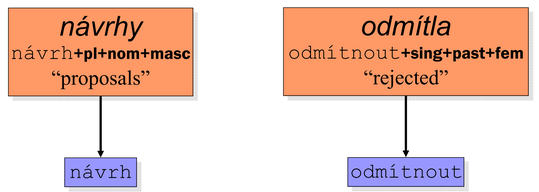
- In our experiments, we used only the person, tense, number, case, and negation tags.

Approach

We experimented with several methods of incorporating morphological analysis into the GIZA++ MT system. In these examples, a box represents a "word." We also show the lemma, simplified morph tags and gloss.

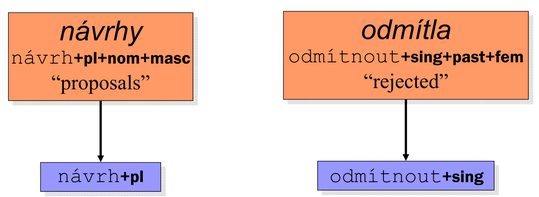
Lemmas

We replace each word with its base (dictionary) form:



Modified Lemmas

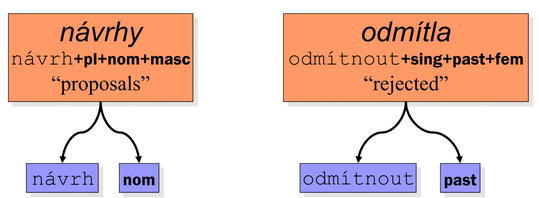
We augment the base form with some morphological tags, discarding the rest:



Retaining **number** tags, as shown here, worked best.

Pseudowords

We split words into separate morphemes to address the fact that some morphemes in Czech correspond to function words in English:



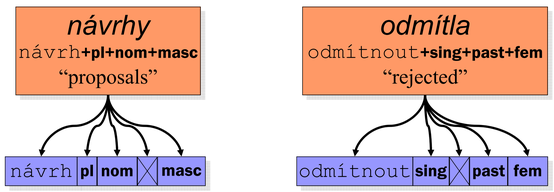
This method worked best for **person** and **tense** tags individually (but surprisingly, not together). Above, we demonstrate the operation on **tense** and **case** tags.

Morphological Alignment Model

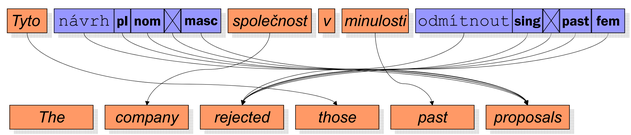
In the previous approaches, we modified the Czech input to the system. Here, we modify the word-to-word alignment model itself to incorporate morphology. In the standard model, the probability of a Czech word f_j aligning to an English word e_i is computed using the expected alignment counts from EM. Here, we compute

$$P(f_j \mid e_i) = \prod_{k=0}^{K} P(f_{jk} \mid e_i)$$

where f_{jk} is the kth morpheme of word f_j.



In this example alignment, two words are analyzed. Each arrow represents an alignment probability.



BLEU scores using the morphological alignment model were not significantly better than those in the modified lemma and pseudoword experiments.

Combined Model

In this experiment, we use pseudowords for the **person** and **negation** tags, and modified lemmas with the **number** and **tense** tags. This method yielded our best results.

Related Work

Niessen and Ney (2000; 2004)

- Found that morphosyntactic analysis and restructuring was helpful for German-English MT on corpus sizes up to 60,000 parallel sentences.

Lee (2004)

- Morphological analysis of POS-tagged English and Arabic text improved MT quality.

Čmejrek et al. (2003)

- Lemmatization of Czech input improved BLEU scores for Czech-English translation.

Al-Onaizan et al. (1999)

- Used morphological and syntactic information from annotated Czech corpus to remove some morphological distinctions and add function words to Czech input, making it more similar to English.
- Subjective measures indicated a slight improvement over lemmatization.

Results

- Using morphological analysis to increase morphosyntactic similarity between input and output languages improves MT quality (measured via the BLEU score)

Method	BLEU
Word-to-word	**0.270**
Truncate all (to 6 characters)	**0.283**
Lemmatize all	**0.299**
Combined morphological model	**0.333**

Reference

S. Niessen and H. Ney. 2004. Statistical machine translation with scarce resources using morphosyntactic analysis. *Computational Linguistics*, 30(2):181–204.

Y. Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings NAACL*.

M. Čmejrek, J. Cuřín, and J. Havelka. 2003. Czech-English dependency-based machine translation. In *Proceedings of EACL*.

Y. Al-Onaizan, J. Cuřín, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. 1999. Statistical machine translation. Final Report, JHU Summer Workshop 1999.

J. Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press.