Statistical Machine Translation
Part VI – Lexical Choice and Morphological Sparsity

Alexander Fraser

CIS, LMU München

2014.11.11 WSD and MT

Lexical problems in SMT

- We are interested in the similarities between WSD and SMT
- This involves lexical choice (~ word-sense choice)
- Phrase-based SMT does a surprisingly good job at this:
  - For instance, there is little POS confusion
  - This is possibly because POS is often selected by nearby words (and these are in the source phrase)
- But there are still significant problems in the output
- Let's look at some error analysis

Vilar et al 2006

- Vilar came up with a taxonomy of errors in SMT output
- I'll present the numbers for an English to Spanish task here

Vilar – Top-level

- Missing words: 19.9% (Filler 12.0%)
- Word order: 15.4%
- Incorrect words: 64.4%
- Unknown words: 0.3%

Vilar – Incorrect Words

- Sense: 21.9%
- Incorrect Form: 33.9% (many verbs)
- Style: 7.9% (repeated words, etc)
- Idioms: 0.7%
- (Extra Words: 0%)

Lexical Features in SMT

- The log-linear framework allows us to combine our different knowledge sources
- The important knowledge sources for lexical choice
  - Phrase-based p(f|e)
  - Phrase-based p(e|f)
  - Lexical p(f|e)
  - Lexical p(e|f)
  - The language model which models target language context

Lexical Probabilities

- Lexical p(f|e)
- Lexical p(e|f)

- Many larger phrases occur just once in the source and target corpora
- Their probabilities at the phrase level are automatically 1
- Using lexical probabilities is a way to "smooth" this

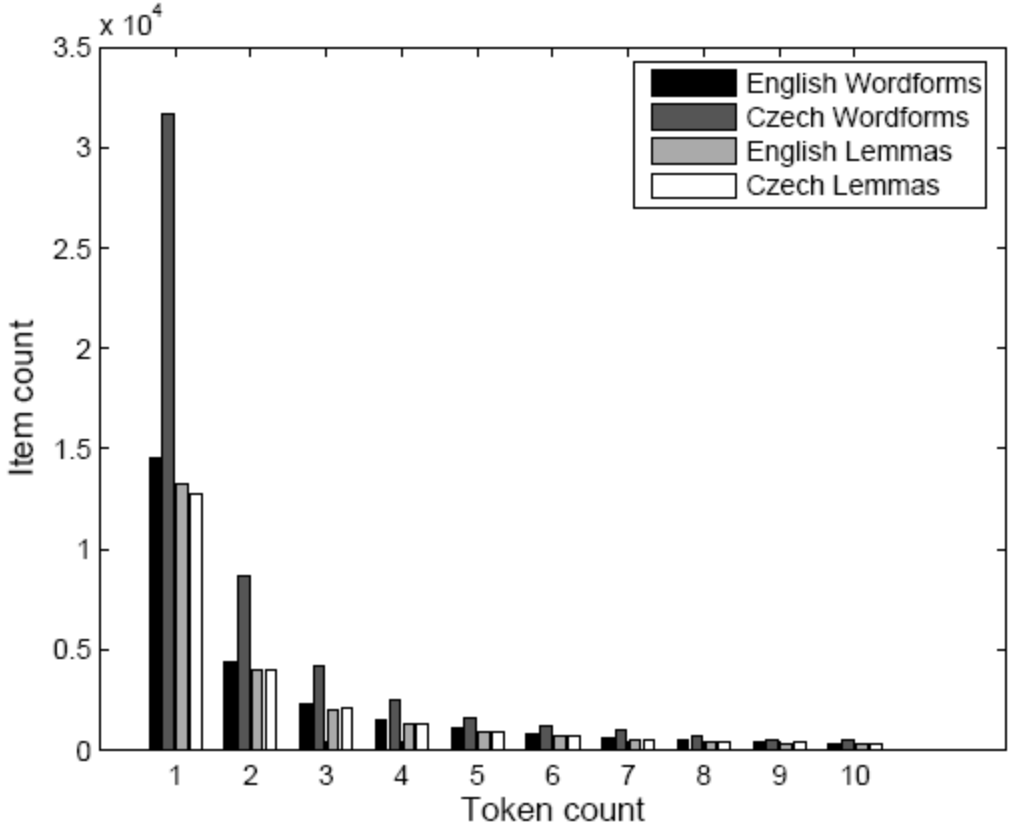
Language Model

- For a somewhat wider context, both source words in the phrase pair used and the language model (capturing target context outside of the phrase pair) are often effective
- However, for some word sense distinctions, this is not enough, we will come back to this later in the seminar, after looking at WSD
- Let's discuss morphology for now

Problems Related to Morphology

- We will use the term morphology loosely here
  - We will discus two main phenomena: Inflection, Compounding
  - There is less work in SMT on modeling of these phenomena than there is on syntactic modeling
    - A lot of work on morphological reduction (e.g., make it like English if the target language is English)
    - Not much work on generating (necessary to translate to, for instance, Slavic languages or Finnish)

Inflection



Goldwater and McClosky 2005

Inflection

- Inflection
  - The best ideas here are to strip redundant morphology
    - For instance case markings that are not used in target language
  - Can also add pseudo-words
    - One interesting paper looks at translating Czech to English (Goldwater and McClosky)
    - Inflection which should be translated to a pronoun is simply replaced by a pseudo-word to match the pronoun in preprocessing

Compounds

- Find the best split by using word frequencies of components (Koehn 2003)
- Aktionsplan -> Akt Ion Plan or Aktion Plan?
  - Since Ion (English: ion) is not frequent, do not pick such a splitting!
  - Until recently not improved by using hand-crafted morphological knowledge
- Now: Fabienne Cap has shown using SMOR (Stuttgart Morphological Analyzer) together with corpus statistics is better (Fritzinger and Fraser WMT 2010)
- This can be taken further by looking at proper names vs. common nouns (e.g., Dinkelacker) and at the (wrong) compositionality assumption behind compounds such as Heckenschütze

- Thanks for your attention!