LING439/539 - Statistical NLP

Philipp Koehn and Kevin Knight (2003). **Empirical Methods for Compound Splitting**. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics.* pp.187-193. PA, USA.

Thursday, September 1 2016

- compounding of words is common in German, Dutch, Finnish, Greek, etc.
- words may be joined freely, this vastly increases the vocabulary size
  - leading to sparse data problems.
- this poses challenges for a number of NLP applications.

Compound splitting

Splitting options for the German word *Aktionsplan* (‘action plan’)

Related work

- Brown (2002) proposed a approach guided by a parallel corpus.
- Monz and de Rijke (2001) and Hedlund et al. (2001) used lexicon based approaches to compound splitting for IR.
- Larson et al. (2000) proposed a data-driven method that combines compound splitting and word recombination for speech recognition.

- Brown (2002) proposed a approach guided by a parallel corpus.
  - The methods leads to improved text coverage of an example based machine translation system
  - no results on translation performance are reported (?)

Brown, R. D. (2002). Corpus-driven splitting of compound words. In *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*.

- Monz and de Rijke (2001) and Hedlund et al. (2001) used lexicon based approaches to compound splitting for IR.
    - stemming does not help the performance of IR systems..
    - splitting compound words will improve results ?

Monz, C. and de Rijke, M. (2001). Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In *Second Workshop of the Cross-Language Evaluation Forum (CLEF)*

Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., and Jarvelin, K. (2001). Utaclir @ CLEF 2001 - effects of compound splitting and n-gram techniques. In*Second Workshop of the Cross-Language Evaluation Forum (CLEF)*

- Larson et al. (2000) proposed a data-driven method that combines compound splitting and word recombination for speech recognition.
    - it reduces the number of out-of-vocabulary words
    - it does not improve speech recognition accuracy.

Larson, M., Willett, D., Kohler, J., and Rigoll, G. (2000). Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *6th International Conference on Spoken Language Processing (ICSLP)*.

Aktionsplan

- aktionsplan
- aktion—plan
- aktions—plan
- akt—ion—plan

aktionsplan, aktions, aktion, akt, ion, and *plan* have been observed as **whole words** in the training corpus.

Recall the example of *Aktionsplan*, where the letter *s* was inserted between *Aktion* and *Plan*. ⇒ **linguistic knowledge**

Frequency Based Metric

Given the count of words in the corpus, we pick the split S with the highest **geometric mean** of word frequencies of its parts p_i (n being the number of parts):

$$\underset{S}{\operatorname{argmax}}\Big(\prod_{p_i \in S} \operatorname{count}(p_i)\Big)^{\frac{1}{n}}$$

Aktionsplan

- **aktionsplan(852) = 852**
- aktion(960)—plan(710) = $(960 * 710)^{1/2} = 825.6$
- aktions(5)—plan(710) = 59.6
- akt(224)—ion(1)—plan(710) = 54.2

Freitag (‘Friday’): *frei* (‘free’) and *Tag* (‘day’):

- **frei(885)—tag(1864) = 1284.4**
- freitag(556) = 556

Guidance from a parallel corpus

Acquisition of splitting knowledge from a parallel corpus: The split *Aktion*—*Plan* is preferred since it has most coverage with the English (two words overlap).