Comparison of Different Modeling Units for Language Model Adaptation for Inflected Languages

Tanel Alumäe
Institute of Cybernetics at Tallinn University of Technology, Estonia

Motivation

Problem: Language model adaptation is a task of building a language model for speech recognition that is better suited for the given domain than a general background model, given a small adaptation corpus. The LSA-based approach gradually adapts the background language model based on the recognized words by boosting the unigram probabilities of semantically related words, using co-occurrence analysis of words and documents. **What kind of units – words, lemmas or morphemes – are best suited for LSA-based LM adaptation for highly inflected languages?**

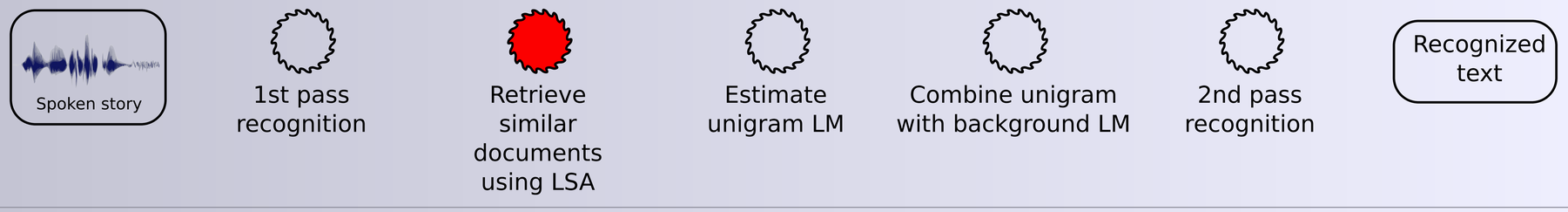
Inflected languages

- Estonian is an agglutinative and inflective language
  - Each word-phrase can occur in a large number of inflected forms
  - Most word forms are formed by adding **suffixes** to **stems**
  - Most words have multiple distinct stems
  - Also, **compound words** are very frequent and can be formed spontaneously
- In LVCSR, morphemes are used as basic language units, to achieve high **coverage**
- Are morphemes good units for LM adaptation, i.e., do morphemes carry enough semantic content?

Latent Semantic Analysis

Latent semantic analysis (LSA) is a technique of analyzing relationships between a set of documents and the terms they contain. LSA uses a term-document matrix which describes the occurrences of terms in documents, using the **bag-of-words** approach. LSA transforms the occurrence matrix into a relation between the terms and some concepts, and a relation between those concepts and the documents. LSA can be used to find semantically related documents, using a handout-text from the given topic.

Architecture



Method

Spoken data is assumed to be segmented into topically homogeneous stories. Each story is recognized as follows:

1. **1st pass recognition**: find the initial recognition hypothesis using a 'background' morpheme-based language model
2. **Retrieve similar documents using LSA**: using the recognized text as a seed, find N semantically closest documents from a large document corpus. Use either lemmas, words or morphemes as basic units in the LSA model.
   (a) Convert the available topic seed to a pseudo-document representation in the LSA space
   (b) Find the vector of the pseudo-document
   (c) Find those documents from the big adaptation corpus that point to the **same direction**
3. **Estimate unigram language model**: use the morpheme unigram statistics in the closest documents to estimate a new unigram language model.
4. **Combine unigram with background language model**:
   - This can be done using *fast marginal adaptation* (FMA)

$$P_{Adap}(w_i|h) = \frac{\alpha(w) P_{BG}(w|h)}{Z(h)}$$

   - $\alpha(w)$ scales word probabilities up or down, depending on their relative frequency in the adaptation corpus with respect to the background corpus:

$$\alpha(w) \approx \left(\frac{P_{Adap}(w)}{P_{BG}(w)} \right)^{\beta}$$

5. **2nd pass recognition**: use the adapted language model to re-recognize the spoken story.

Experiments

- LSA model:
  - $\sim$ 500 000 documents (mainly newspaper articles) were used for building LSA models
  - 3 different models were built: based on words, lemmas and morphemes
  - For all models, 60 000 most frequent units were used
  - Coverage:
    - Words: 86%
    - Lemmas: 94%
    - Morphemes: 98%
- Test data: hourly short broadcast news recordings from the Estonian national radio, manually segmented into stories and sentences
  - 44 stories, 193 utterances
- LM was adapted seperately to each story
- We measured letter error rate (LER) without and with adaptation
- Morpheme-based adaptation statistically significantly better

System	LER, %	Change
No adaptation	7.1	Baseline
Word-based adaptation	6.7	-6%
Lemma-based adaptation	6.6	-7%
Morpheme-based adaptation	**6.4**	**-9%**

Conclusion

- We experimented with different kinds of language units for Estonian LM adaptation, using LSA for document similarity detection
- Morphemes worked the best while plain words worked the worst
- Applicable to other inflected languages