Language Modeling for Speech Recognition in Agglutinative Languages

Ebru Arısoy Murat Saraçlar

September 13, 2007

Outline

- Agglutinative languages
  - Main characteristics
  - Challenges in terms of Automatic Speech Recognition (ASR)
- Sub-word language language modeling units
- Our approaches
  - Lattice Rescoring/Extension
  - Lexical form units
- Experiments and Results
- Conclusion
- Ongoing Research at OGI
- Demonstration videos

Agglutinative Languages

- Main characteristic: Many new words can be derived from a single stem by addition of suffixes to it one after another.

- Examples: Turkish, Finnish, Estonian, Hungarian...

  Concatenative morphology (in Turkish):

  ∗ nominal inflection: ev+im+de+ki+ler+den

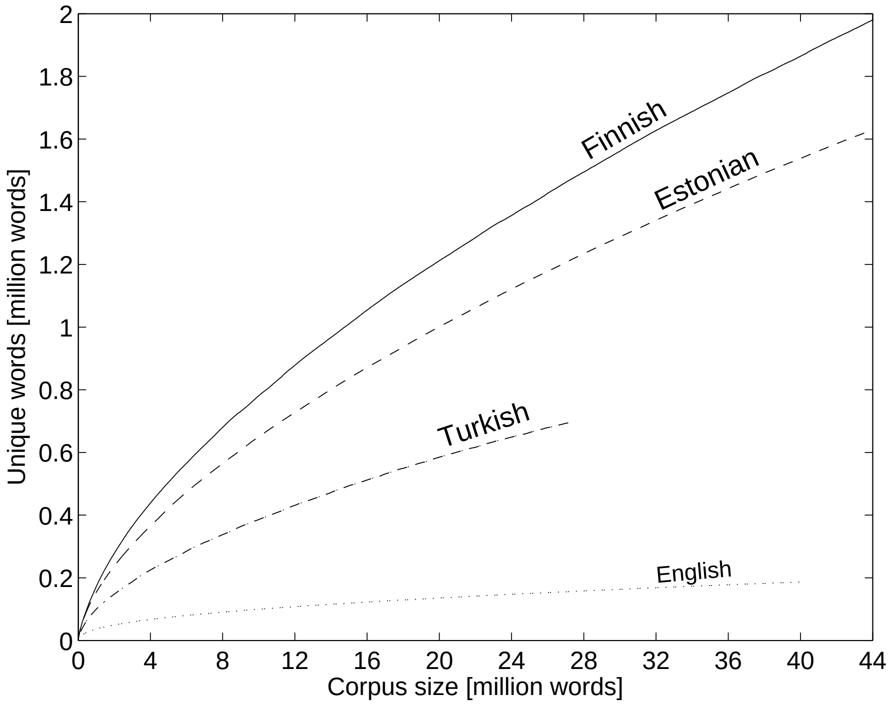
  (one of those that were in my house)

  ∗ verbal inflection: yap+tır+ma+yabil+iyor+du+k

  (It was possible that we did not make someone do it)

- Other characteristics: Free word order, Vowel harmony

Agglutinative Languages – Challenges for LVCSR (Vocabulary Explosion)



- Moderate vocabulary (50K) results in OOV words.
- Huge vocabulary (>200K) suffers from non-robust language model estimates. (Thanks to Mathias Creutz for the Figure)

Agglutinative Languages – Challenges for LVCSR (Free Word Order)

- The order of constitutes can be changed without affecting the grammaticality of the sentence.

  Examples (in Turkish):

  - The most common order is the SOV type (Erguvanlı, 1979).
  - The word which will be emphasized is placed just before the verb (Oflazer and Bozşahin, 1994).

    Ben çocuğa **kitabi** verdim *(I gave the book to the children)*

    Çocuga kitabi **ben** verdim *(It was me who gave the child the book)*

    Ben kitabi **çocuga** verdim *(It was the child to whom I gave the book)*

  Challenges:

  - Free word order causes "sparse data".
  - Sparse data results in "non-robust" N-gram estimates.

Agglutinative Languages – Challenges for LVCSR (Vowel Harmony)

- The first vowel of the morpheme must be compatible with the last vowel of the stem.

  Examples(in Turkish):

  – Stem ending with back/front vowel takes a suffix starting with back/front vowel.

    ✓ağaç+lar *(trees)* ✓çiçek+ler *(flowers)*

  – There are some exceptions: ✘ ampul+ler *(lamps)*

  Challenges:

  – No problem with words !!!

  – If sub-words are used as language modeling units:

    ∗ Words will be generated from sub-word sequences.

    ∗ Sub-word sequences may result in ugrammatical items

Words vs. Sub-words

- Using words as language modeling units:
  - ✘ Vocabulary growth – > Higher OOV rates.
  - ✘ Data sparseness – > non-robust language model estimates.

- Using sub-words as language modeling units:

  (Sub-words must be "meaningful units" for ASR !!!)

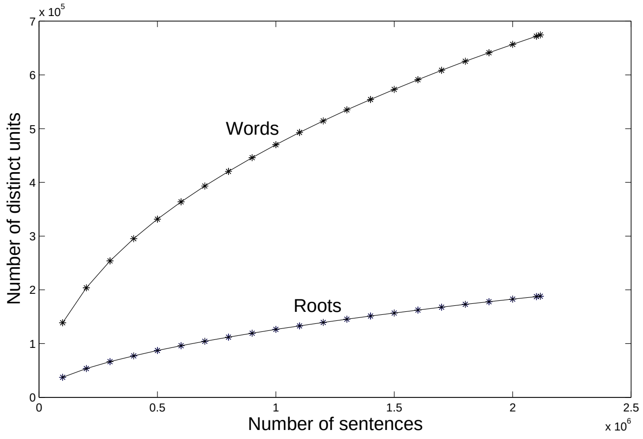
  - ✓ Handle OOV problem.
  - ✓ Handle data sparseness.
  - ✘ Results in ungrammatical, over generated items.

Our Research

- Our Aim:
  - To handle "data sparseness"
    - Root-based models
    - Class-based models
  - To handle "OOV words"
    - Vocabulary extension for words
    - Sub-words recognition units
  - To handle "over generation" by sub-word approaches
    - Vocabulary extension for sub-words
    - Lexical sub-word models

Modifications to Word-based Model

(Arisoy and Saraclar, 2006)



- Root-based Language Models

  Main idea: Roots can capture regularities better than words

  $P(w_3|w_2, w_1) \approx P(r(w_3)|r(w_2), r(w_1))$

- Class-based Language Models

  Main idea: To handle data sparseness by grouping words

  $P(w_3|w_2, w_1) = P(w_3|r(w_3)) * P(r(w_3)|r(w_2), r(w_1))$

Modifications to Word-based Model

(Arisoy and Saraclar, 2006)

- Vocabulary Extension (Geutner et al., 1998)

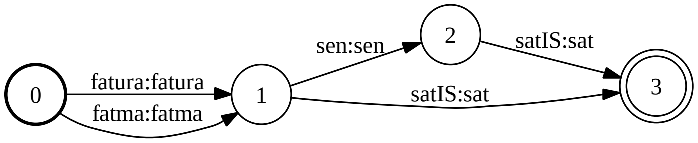
  Main idea: To extend the utterance lattice with similar words, then perform second pass recognition with a larger vocabulary language model

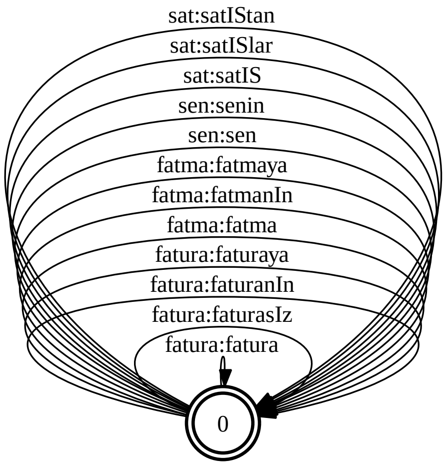
  – Similarity criterion: "having the same root"

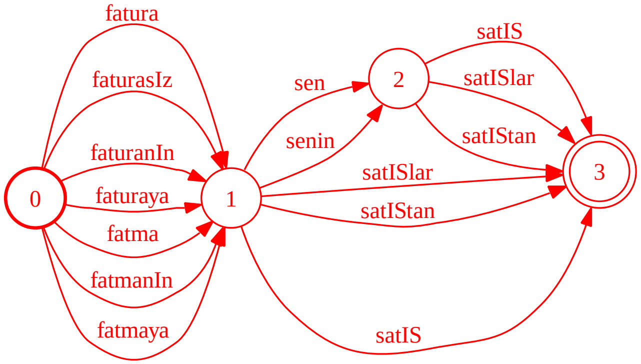
  – A single language model is generated using all the types (683K) in the training corpus.

Modifications to Word-based Model

- Vocabulary Extension







Sub-Word Approaches (Background)

- Morpheme model:
  - Require linguistic knowledge (Morphological analyzer)

    Morphemes: kes il di ği # an dan # itibaren

- Stem-ending model:
  - Require linguistic knowledge (Morphological analyzer, stemmer)

    Stem-endings: kes ildiği # an dan # itibaren

Sub-Word Approaches (Background)

- Statistical morph model (Creutz and Lagus, 2005):
  - Main idea: To find an optimal encoding of the data with concise lexicon and the concise representation of corpus.
    - ∗ Unsupervised
    - ∗ Data-driven
    - ∗ Minimum Description Length (MDL)

Morphemes: kes il di ği # an dan # itibaren

Morphs: kesil diği # a ndan # itibar en

Sub-Word Approaches

- Statistical Morph model is used as the sub-word approach.
  - Dynamic vocabulary extension is applied to handle ungrammatical items.

- Lexical stem ending models are proposed as a novel approach.
  - Lexical to surface form mapping ensures correct surface form alternations.

Modifications to Morph-based Model

(Arisoy and Saraclar, 2006)

- Vocabulary Extension

  Motivation:

  – 159 morph sequences out of 6759 do not occur in the fallback (683K) lexicon. Only 19 are correct Turkish words.

  – Common Errors: Wrong word boundary, incorrect morphotactics, meaningless sequences

  – Simply removing non-lexical arcs from the lattice increases WER by 1.8%.

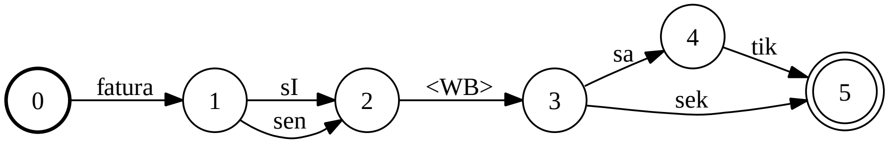
  Main idea: To remove non-vocabulary items with a mapping from morph sequences to grammatically correct similar words, then perform second pass recognition.

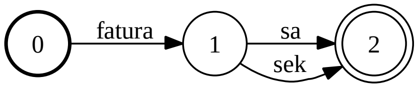
  – Similarity criterion is "having the same first morph"

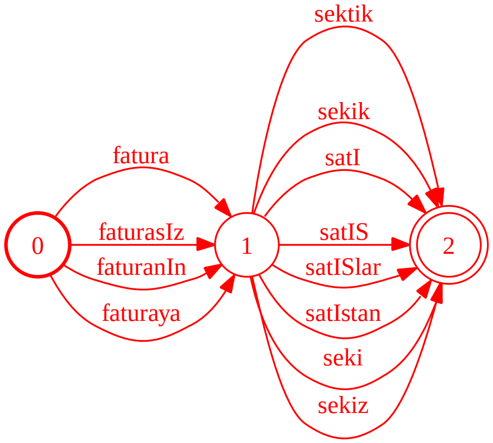
Modifications to Morph-based Model

(Arisoy and Saraclar, 2006)

- Vocabulary Extension







Lexical Stem-ending Model (Arisoy et al., 2007)

Motivation:

- Same stems and morphemes in lexical form may have different phonetic realizations

Surface form:	`ev-ler` (houses)	`kitap-lar` (books)
Lexical from:	`ev-lAr`	`kitap-lAr`

Advantages:

- Lexical forms capture the suffixation process better.
- In lexical to surface mapping;
  - compatibility of vowels is enforced.
  - correct morphophonemic is enforced regardless of morphotactics.

Comparison of Language Modeling Units

	Lexicon Size	Word OOV Rate (%)
Words	50K	9.3
Morphs	34.7K	0
Stem-endings		
Surf:	50K (40.4K roots)	2.5
Lex:	50K (45.0K roots)	2.2

Experiments and Results

- Newspaper Content Transcription
  - Baseline Word and Morph System
  - Lattice re-scoring with root-based and class-based models for word baseline.
  - Dynamic Vocabulary extension for word and morph baselines.

- Broadcast News (BN) Transcription
  - Broadcast News database is collected.
  - Various sub-word approaches are investigated.
  - BN transcription and retrieval systems are developed (Demonstration videos will be shown)

Experimental Setup
(Newspaper Content Transcription)

- Text corpus [a]: 26.6M words
- Acoustic Train Data: 17 hours of speech – 250 speakers
- Test Data: 1 hour of newspaper sentences – 1 female speaker
- Language Modelling: SRILM (Stolcke, 2002) toolkit with interpolated modified Kneser-Ney smoothing
- Decoder [b]: AT&T Decoder (Mohri and Riley, 2002)

[a] Thanks to Sabanci and ODTU universities for text and acoustic data
[b] Thanks to AT&T Labs–Research for the software

Baseline systems
(Newspaper Content Transcription)

Baseline Language Models: 3-gram **(words)** and 5-gram **(morphs)**

Experiments	Lexicon	OOV	WER	LER
		Test (%)	(%)	(%)
Baseline-word	50K	11.8	38.8	15.2
Baseline-word	120K	5.6	36.0	14.1
Baseline-morph	34.3K	0	33.9	12.4
Baseline-word (cheating)	50.7K	0	30.0	11.9

Results

Rescoring Experiments:

– Original (word) and new (root, class) language models are interpolated with an interpolation constant.

– Lattice rescoring strategy is applied.

✓ Root-based: 38.8% → 38.3% (0.5% absolute reduction)

✘ Class-based: 38.8% (baseline)

Results

Vocabulary Extension Experiments:

– Original (word/morph) lattice is extended with new words from the full lexicon using root/first-morph similarity.

– Second pass recognition is performed with the full word vocabulary language model.

Unit	Experiment	WER	LER	LWER
Word	Baseline (50K)	38.8	15.2	15.5
	Extended Lattice	36.6	14.3	9.6
Morph	Baseline (34.3K)	33.9	12.4	14.7
	Extended Lattice	32.8	12.2	6.0

Experimental Setup

(Broadcast News (BN) Transcription)

- Text corpus [a]: 96.4M words
- Acoustic Train Data: 68.6 hours of BN from 6 different channels
- Test Data: 2.4 hours of BN from 5 different channels
- Language Modelling: SRILM (Stolcke, 2002) toolkit with interpolated modified Kneser-Ney smoothing
- Decoder [b]: AT&T Decoder (Mohri and Riley, 2002)

[a] Thanks to Sabanci and ODTU universities for text data

[b] Thanks to AT&T Labs–Research for the software

Experimental Setup
(Broadcast News (BN) Transcription)

Breakdown of data in terms of acoustic conditions (in hours)

Partition	f0	f1	f2	f3	f4	fx	Toplam
Training	25.9	7.0	1.8	6.2	26.4	1.3	68.6
Test	1.27	0.11	0.10	0.20	0.83	0.03	2.54

f0: clean
f1:spontaneous
f2:telefon speech
f3:music background
f4:degraded acoustic conditions
f5:non-native speaker
fx:other

Experiments

1. Baseline Models:
   - Same acoustic model and unit specific language models are used.
   - The size of the language models is set with entropy-based pruning (Stolcke, 1998).
2. Re-scoring strategy:
   - Lattice output of the recognizer is re-scored with a same order n-gram language model pruned with a smaller pruning constant.
   - Only applied to sub-word units.
3. Channel Adapted Acoustic Models:
   - Acoustic models are adapted for each channel. (Supervised MAP adaptation)

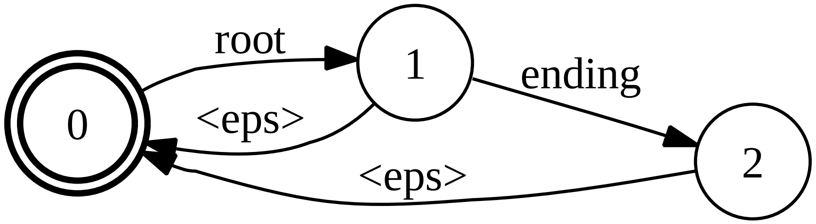
Experiments

4. Restriction:

   – Applied to stem ending models.

   – Aim is to enforce the decoder not to generate consecutive ending sequences.

   – This restriction is implemented as a finite state acceptor that is intersected with the lattices.



Results

Experiments	f0	Avg.
Words	27.7	41.4
Morphs_rescore	22.4	37.9
Stem-ending_rescore	24.7	38.8
Stem-ending-lexical_rescore	21.1	37.0
Words_map_sup	26.3	39.6
Morphs_map_sup_rescore	19.9	35.4
Stem-ending_map_sup_rescore	23.1	36.5
Stem-ending-lexical_map_sup_rescore	19.4	34.6

f0: Clean speech

Conclusion

- Newspaper Content Transcription
  - Baseline word-model: 38.8%
    - ✓ Root-based model 38.8% → 38.3% (0.5% reduction)
    - ✘ Class-based model
    - ✓ Dynamic vocabulary extension 38.8% → 36.6% (2.2%)
  - Baseline morph-model: 33.9%
    - ✓ Dynamic vocabulary extension 33.9% → 32.8% (1.1%)
- Broadcast News Transcription
  - ✓ Sub-word approaches perform better than words.
  - ✓ Lexical stem-ending model significantly improves WER by 0.8% over the previous best model using statistical morphs.

Ongoing Research – 1

- Broadcast News Transcription System is built with IBM tools.

Experiments	f0	f1	f2	f3	f4	fx	Avg.
Test							
CD	23.8	43.0	39.3	32.8	44.2	34.3	33.1
VTLN	23.1	42.2	37.5	29.8	41.5	33.8	31.4
FSA-SAT (SI)	22.5	37.4	36.5	28.0	38.9	28.7	29.9
FSA-SAT (SD)	22.4	36.0	31.4	27.5	38.4	28.2	29.2

Ongoing Research – 2

- Discriminative Language Modeling (DLM) for Turkish
  - How to generate the training data for DLM?
    - Effect of over-trained language models
    - Effect of over-trained acoustic models
  - What are the discriminative features for Turkish?
    - Word n-grams (decreases WER approximately 0.6%)
    - Morphological features
    - Syntactic features

Acknowledgements

We would like the thank Hasim Sak for his contribution to lexical stem-ending models.

We would like to thank Siddika Parlak and Ismail Ari for preparing the BN retrieval demonstration.

References

Arisoy, E., Sak, H., Saraclar, M., 2007. Language modeling for automatic Turkish broadcast news transcription. In: Interspeech-Eurospeech 2007. Antwerp, Belgium.

Arisoy, E., Saraclar, M., 2006. Lattice extension and rescoring based approaches for LVCSR of Turkish. In: nternational Conference on Spoken Language Processing - Interspeech2006 ICSLP. Pittsburg PA, USA.

Creutz, M., Lagus, K., 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March.

Erguvanlı, E., 1979. The function of word order in Turkish grammar. Ph.D. thesis, University of California, Los Angeles, USA.

Geutner, P., Finke, M., Scheytt, P., Waibel, A., Wactlar, H., 1998. Transcribing multilingual broadcast news using hypothesis driven lexical adaptation. In: DARPA Broadcast News Workshop. Herndon, USA.

Mohri, M., Riley, M. D., 2002. Dcd library, speech recognition decoder library, AT&T Labs - Research. http://www.research.att.com/sw/tools/dcd/.

Oflazer, K., Bozşahin, H. C., 1994. Turkish natural language processing initiative: An overview. In: Proceedings of the Third Turkish Symposium on Artificial Intelligence and Artificial Neural Networks. Ankara, Turkey.

Stolcke, A., 1998. Entropy-based pruning of backoff language models. In: Proc. DARPA Broadcast News Transcription and Understanding Workshop. Lansdowne, VA, pp. 270–274.

Stolcke, A., 2002. Srilm – An extensible language modeling toolkit. In: Proc. ICSLP 2002. Vol. 2. Denver, pp. 901–904.

Questions???