Analyzing Stemming Approaches for Turkish Multi-Document Summarization

Muhammed Yavuz Nuzumlalı and Arzucan Özgür
{yavuz.nuzumlali, arzucan.ozgur}@boun.edu.tr
Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

Motivation

- Automatic MDS enables to extract the most valuable information from a set of documents about the same topic in a condensed form.
- There are limited number of studies about MDS for morphologically rich languages.
- Previous studies on other IR problems show that applying morphological analysis may improve performance for Turkish.

Turkish Morphology

- Agglutinative
- Roots can take one or more inflectional and derivational affixes.
- # of unique terms in Turkish is three times more than English for a corpus of 1M words.

Problems:

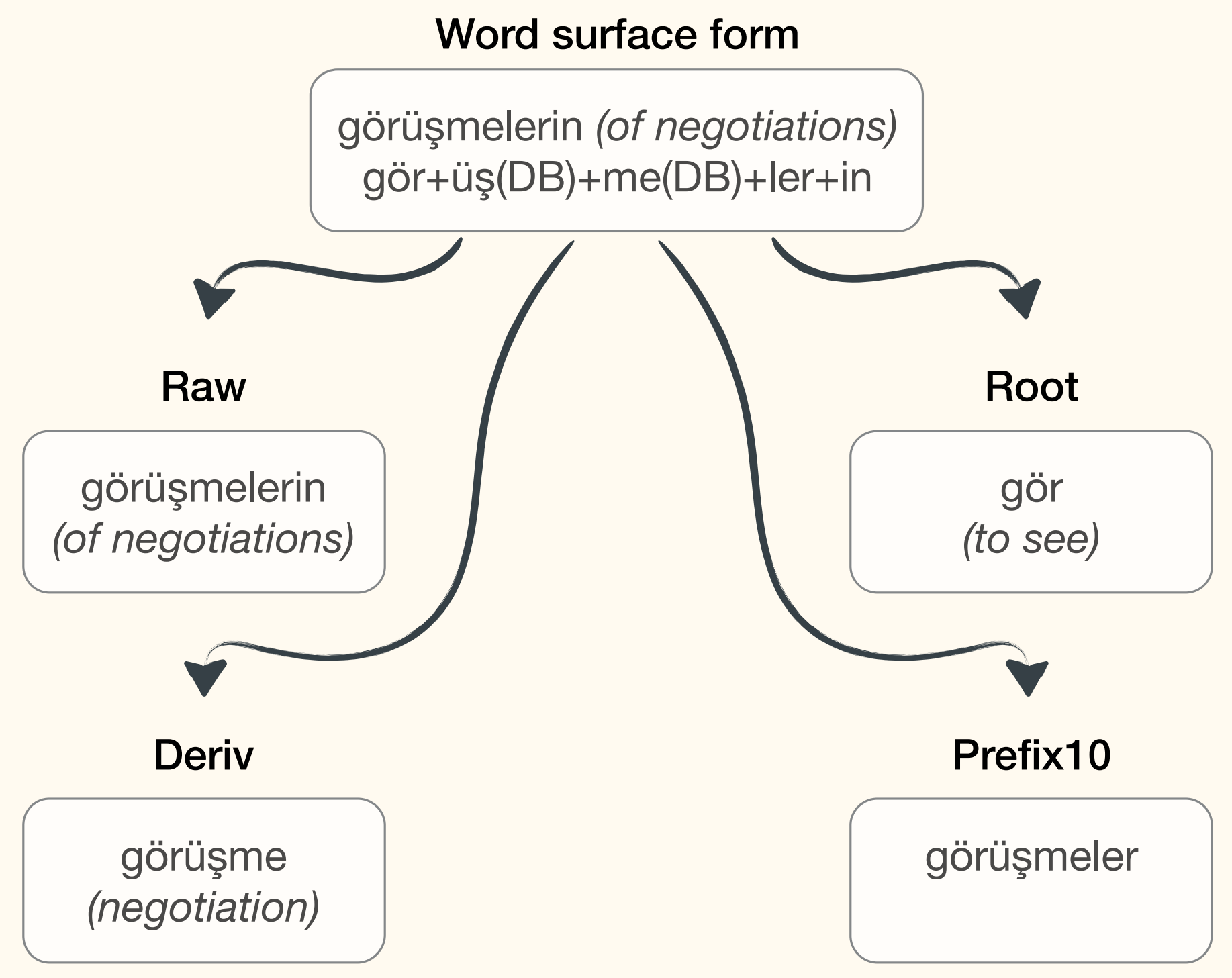
- Data sparseness
- Morphological ambiguity

Word	Analysis
gören *(the one who sees)*	gör+en(DB)
görülen *(the one which is seen)*	gör+ül(DB)+en(DB)
görüş *(opinion)*	gör+üş(DB)
görüşün (your opinion)	gör+üş(DB)+ün
görüşler *(opinions)*	gör+üş(DB)+ler
görüşme *(negotiation)*	gör+üş(DB)+me(DB)
görüşmelerin *(of negotiations)*	gör+üş(DB)+me(DB)+ler+in

Stemming Policies

Methods:

- **Raw :** Take the surface forms w/o modification.
- **Root :** Take the most simple unit, the root.
  - May cause oversimplification!
- **Deriv :** Discard only inflectional affixes.
  - Solves oversimplification issue.
- **Prefix :** Take the first n letters (n = threshold).
  - In Turkish, affixes almost always occur as suffixes.
  - Simple and fast.



Word surface form

görüşmelerin *(of negotiations)*
gör+üş(DB)+me(DB)+ler+in

Raw

görüşmelerin
(of negotiations)

Root

gör
(to see)

Deriv

görüşme
(negotiation)

Prefix10

görüşmeler

- We used a two-level morphology analyzer (Oflazer, 1994) and a perceptron-based morphological disambiguator (Sak et. al., 2007).
  - Root and Deriv forms are generated from disambiguator output.

LexRank

- Graph-based. Challenging baseline for MDS. (Erkan and Radev, 2004)
- Connectivity graph:
  - Nodes: sentences
  - Edges: cosine similarities
- Uses PageRank to find most important sentences.

Data Set

- Created from scratch.
- Tried to mimic DUC 2004 standards.
- 21 topic clusters collected from news domain, each having approximately 10 documents.
  - 337 words per document.
  - 6.84 letters per word.
- Human summaries don't exceed 120 words.
- Annotated by 3 annotators.
- Available @github!
  - https://github.com/manuyavuz/TurkishMDSDataSet_alpha

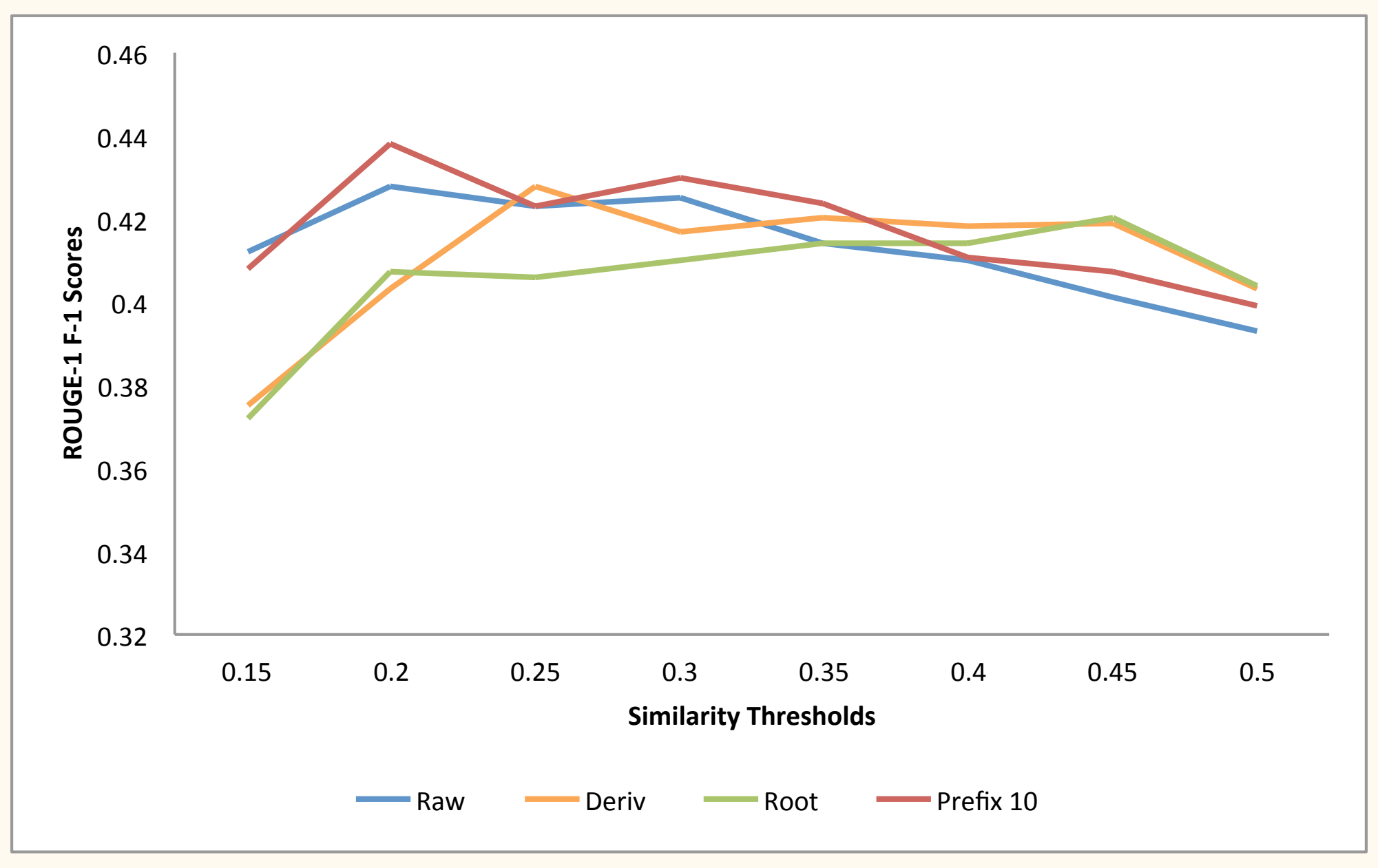
Results

ROUGE Scores:

Policy	ROUGE-1	ROUGE-2	ROUGE-W
Prefix10	**0,438**	0,194	**0,197**
Prefix12	0,433	**0,197**	0,195
Prefix9	0,432	0,194	0,194
Prefix4	0,432	0,178	0,190
Prefix7	0,431	0,189	0,190
Prefix5	0,431	0,183	0,190
Prefix6	0,430	0,185	0,189
Raw	0,428	0,189	0,191
Deriv	0,428	0,178	0,188
Prefix8	0,427	0,187	0,188
Prefix11	0,427	0,190	0,193
Root	0,420	0,180	0,180

- Prefix outperforms Raw. (Prefix10 is best).
- Deriv performs similar with Raw.
- Root is worst.

Effect of Similarity Threshold:



- Used during sentence selection process.
  - Do NOT select the sentence if it's very similar to previously selected sentences.
- Root gets best score when threshold is high.
- Others gets best score when threshold is low.

Discussion

- Bad performance of Root is expected.
  - We lose semantic differences provided by derivational affixes.
- To analyze result of Deriv, we used an entropy-based measure.

$$D_{Deriv_i} = \{t \mid t\, inflected\ from\ Deriv\ i\}$$
$$H(Deriv_i) = \sum_{t \in D_{Deriv_i}} -p(t) \log p(t)$$
$$H(C) = \sum_i \frac{H(Deriv_i)}{N}$$

  - Helps to quantify homogeneity of clusters.
- Consider the deriv "görüşme" *(negotiation)*
  - Assume that it occurs 8 times in two different documents with the following distribution:

Surface Form	Doc1	Doc2
görüşmede *(on negotiation)*	2	2
görüşmeler *(negotiations)*	4	6
görüşmenin *(of negotiation)*	2	0
H(görüşme)	**1,5**	**0,81**

  - Documents having lower entropy value are more homogenous.
- Generate random clusters to compare with topic clusters.
  - Randomly select 10 different clusters.
  - Randomly select 1 document from each selected cluster.
- Avg. entropy of Topic Clusters (Data Set) : **4,99**
- Avg. entropy of Random Clusters : **7,58**
  - Statistically significant (p = 0,05)

Hypothesis:

- Topic clusters are more homogenous.
- Deriv forms are usually seen in the same surface form among documents in a topic cluster.
- Therefore, applying Deriv does NOT affect performance much.

Conclusion

- Fixed-length truncation methods improves scores.
- Surprisingly, morphological analysis does not improve performance.
  - Possibly due to homogeneousness of the documents in a cluster.

Future work:

- Apply sentence simplification methods.
- Extend data set with more reference summaries and more topic clusters.

References

- Kemal Oflazer. 1994. Two-level description of turkish morphology. *Literary and linguistic computing*, 9(2):137–148.
- Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological disambiguation of turkish text with perceptron algorithm. In Alexander F. Gelbukh, editor, *CICLing*, volume 4394 of *Lecture Notes in Computer Science*, pages 107–118. Springer.
- Güneş¸ Erkan and Dragomir R. Radev. 2004. Lex-pagerank: Prestige in multi-document text summarization. In *EMNLP*, pages 365–371. ACL.

EMNLP 2014: Conference on Empirical Methods in Natural Language Processing, Doha, Qatar.