



Story Link Detection in Turkish Corpus

A. Can Polatkan, Güven Köse, Hamid Ahmadi, Yaşar Tonta

polatkan@informatik.uni-tuebingen.de

guvenkose@gmail.com

hamid2026@gmail.com

tonta@hacettepe.edu.tr

University of Tübingen

Center for Bioinformatics
Integrative Transcriptomics

Hacettepe University

Department of Information Management

11:30 - 11:45

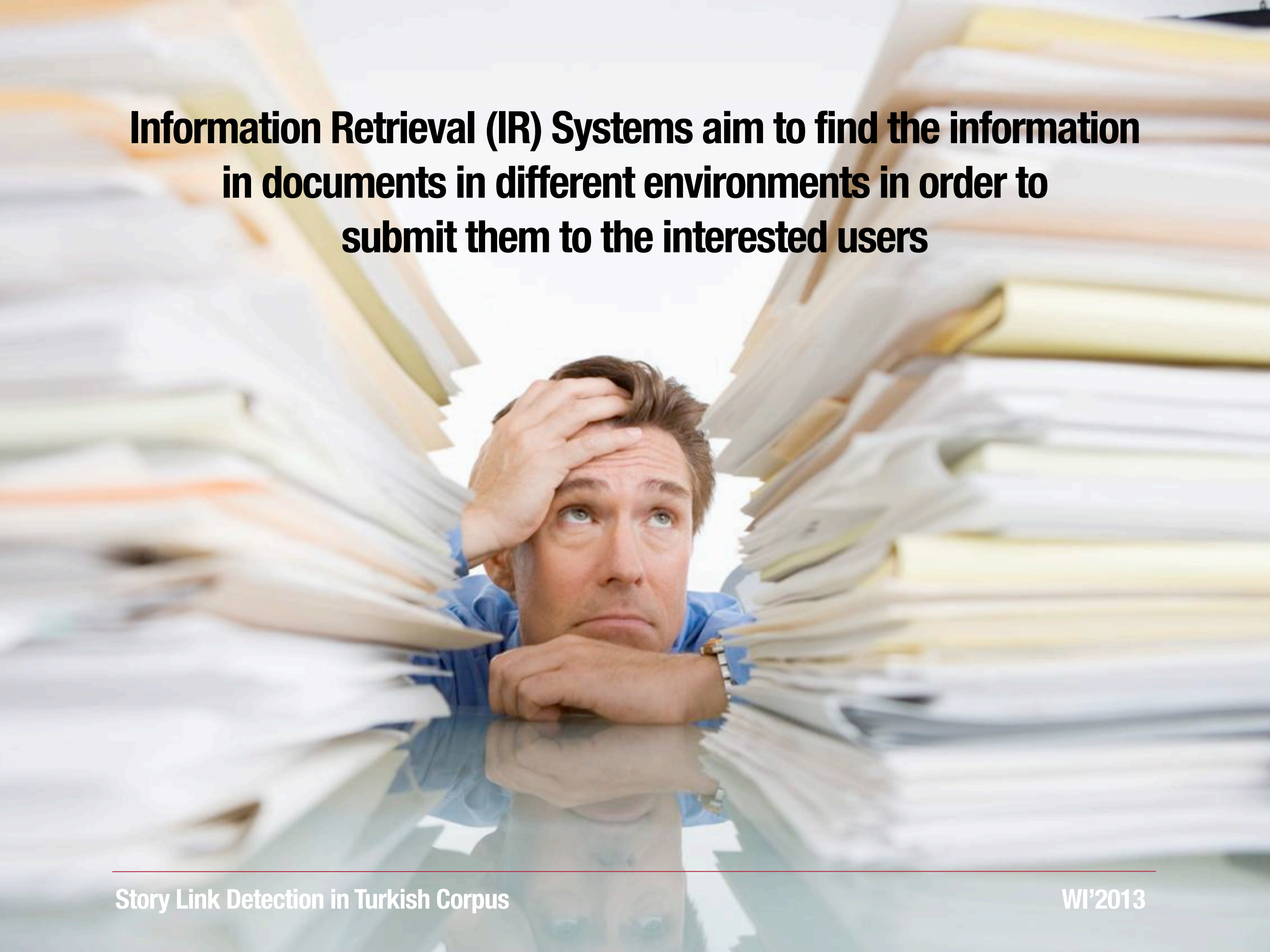
Tower Room 1204

Session 6: Web Data Analysis



Motivation

**Information Retrieval (IR) Systems aim to find the information
in documents in different environments in order to
submit them to the interested users**





Motivation

*In the recent years IR systems mainly focus on
Topic Detection and Tracking (TDT)*

- ▶ TDT aims to
 - ▶ detect new unreported stories
 - ▶ organize them temporally
 - ▶ link incoming news items with previously detected stories on the same topic
 - ▶ monitor news streams online till stories peter out



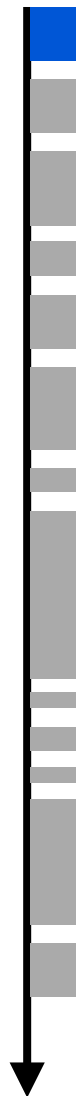
Motivation

- ▶ TDT try to solve the tasks below:
 - ▶ First Story Detection
 - ▶ Story Clustering
 - ▶ Topic Tracking
 - ▶ Story Link Detection

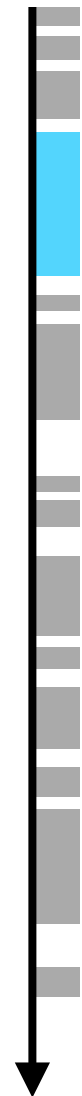


First Story Detection

AP



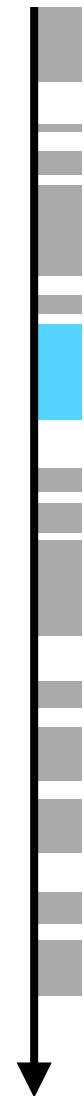
**BBC
NEWS**



CNN



DW



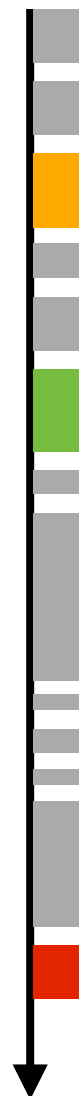
find the first story
about the topic





Story Clustering

AP



**BBC
NEWS**



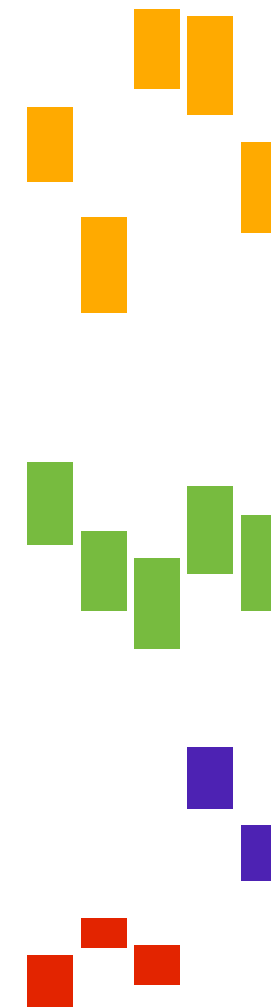
CNN



DW



put stories of the
same topic together





Topic Tracking

AP



**BBC
NEWS**



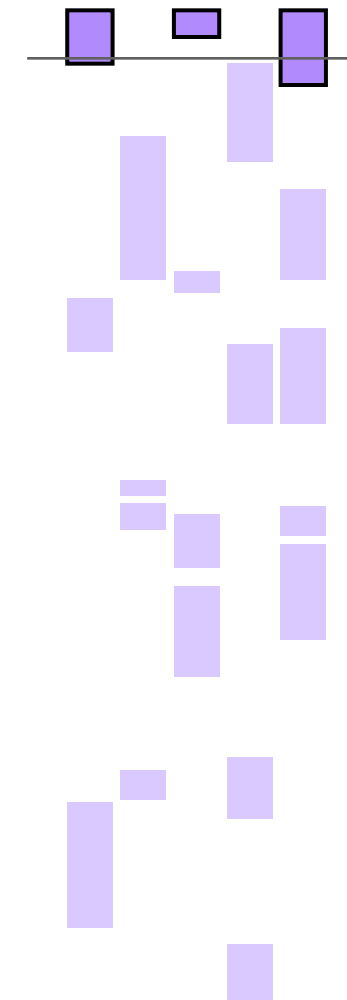
CNN



DW

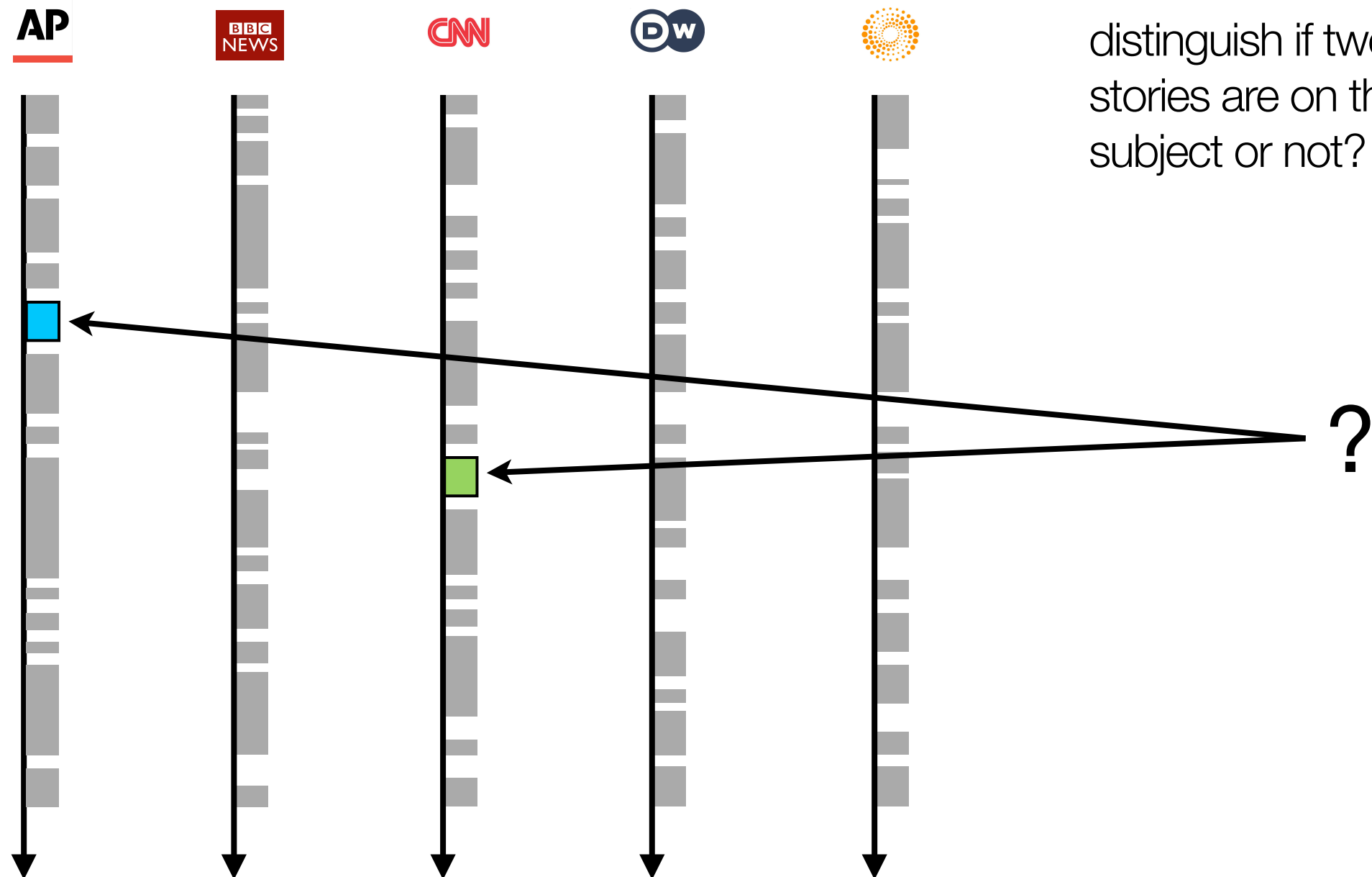


given a few stories on
the topic, find the rest





Story Link Detection



Olympic Torch Lands Back on Earth After Spacewalk

1 hour ago from **T**ime and 315 more



It was launched Thursday as part of a Russian stunt before the 2014 Winter Olympics

Like TIME on Facebook for more breaking news and current events from around the globe!



source: news360.com

Olympic Torch Lands Back on Earth After Spacewalk


1 hour ago from  Time

315 MORE



9 hours ago from  BBC NEWS


Sochi Olympic torch returns to Earth after spacewalk

8 hours ago from  NASA

A warm welcome for the crew

9 hours ago from  NASA

Expedition 37 crew back on Earth

11 hours ago from  CBS News

Welcome home: Soyuz spacecraft, with 3 crew, returns safely to Earth

Thursday as part of a Russian stunt
Winter Olympics

bookmark for more breaking news and
om around the globe!



source: news360.com

< Prev

Next >



Motivation



IR T-IR TDT SLD

- ▶ Story link detection (SLD) algorithms play a key role in establishing linkages between stories discussing the same subject
- ▶ Carrying out the SLD task successfully is expected to solve many problems in TDT

Allan et al.



Methodology



TDT Test Collection

*We performed a study on **Turkish** news items*

- ▶ BilCol-2005 test collection contains 209,305 items
 - ▶ 5 different Turkish news sources
 - ▶ 5,882 news items classified under 80 topics
 - ▶ 203,423 items classified as unknown



Retrieval Models used in the study

Retrieval models used in SLD are similar to traditional IR systems

- ▶ Boolean Model
- ▶ Vector Space Model
- ▶ Probabilistic Model
- ▶ Language Model
- ▶ Relevance Model

In this study, we used the Vector Space Model (VSM) and Relevance Model (RM) to carry the SLD tasks



Retrieval Evaluation

- ▶ We assessed performance by computing
 - ▶ Precision
 - ▶ Recall
 - ▶ F-measure

Assumption:

high precision, recall & f-measure = Better Results



Vector Space Model

- ▶ Documents and queries are represented as vectors of index terms
- ▶ Similarity between these vectors prove the document/query matchup
- ▶ Coefficients contained in the vectors highlight the importance of each term to what extent it represents the documents and/or queries



Vector Space Model

- ▶ Represent the query as a weighted tf.idf vector
- ▶ Represent each document as a weighted tf.idf vector
- ▶ Compute the cosine similarity score for the query vector and each document vector
- ▶ Rank documents with respect to the query by score



Relevance Model

- ▶ RM is the advanced version of the language model which is extensively used in SLD tasks
- ▶ RM offers a new approach to the estimation of probabilities when the necessary conditions of training data are absent
- ▶ The probability distributions were compared on the basis of Kullback-Leibler to determine document similarity



Relevance Model

- ▶ Represent the query as a probability distribution
- ▶ Represent each document as a probability distribution
- ▶ Compute the Kullback-Leibner divergence score for the query and each document
- ▶ Rank documents with respect to the query by score



Testing



Test Collection

The BilCol-2005 collection is divided into training (1/3 of news items) and test (2/3 of news items) sets

- ▶ Tests carried through the Turkish corpus with
 - ▶ 3,922 news items with known topic titles
 - ▶ 135,609 news items with unknown topic titles



TDT Test Collection

- In order to identify the effects of the number of index terms on the match performance, tests repeated for 1, 2, 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 225, 250, 275, 300, 400, 500 and 1000 respectively

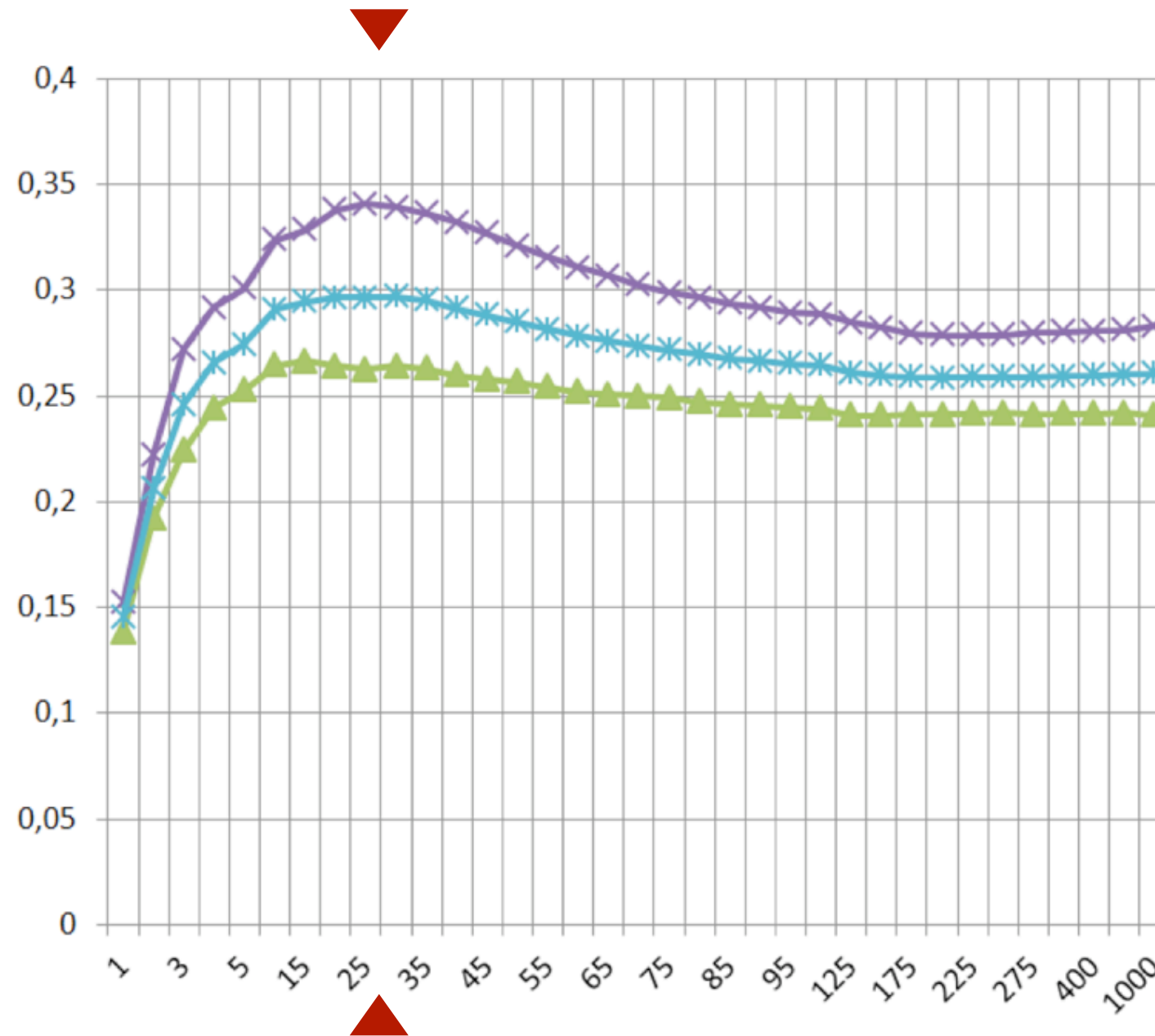
1 5 10 25 50 100 175 250 500 1 000



Discussion and Conclusion

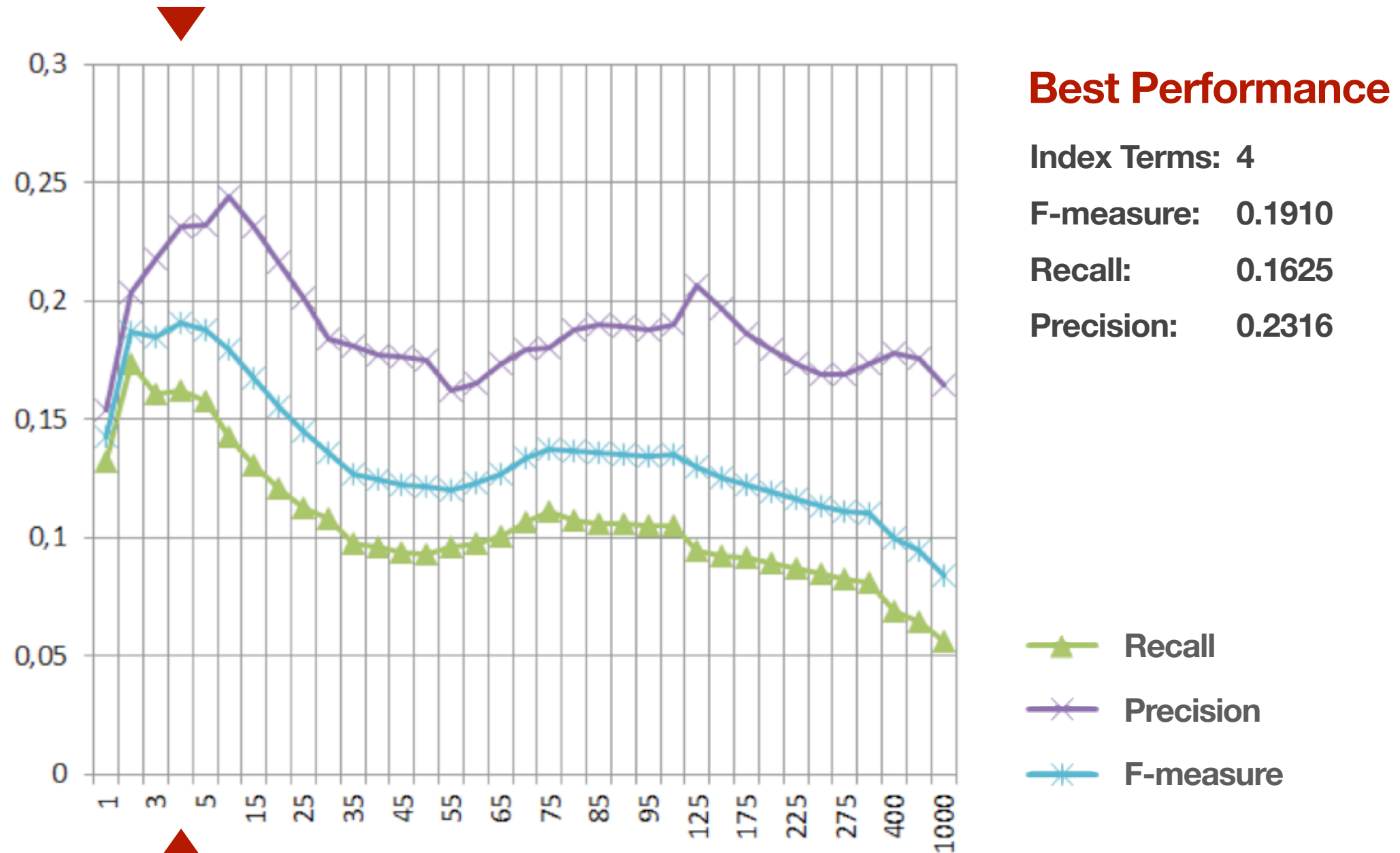


Test Results of Vector Space Model





Test Results of Relevance Model

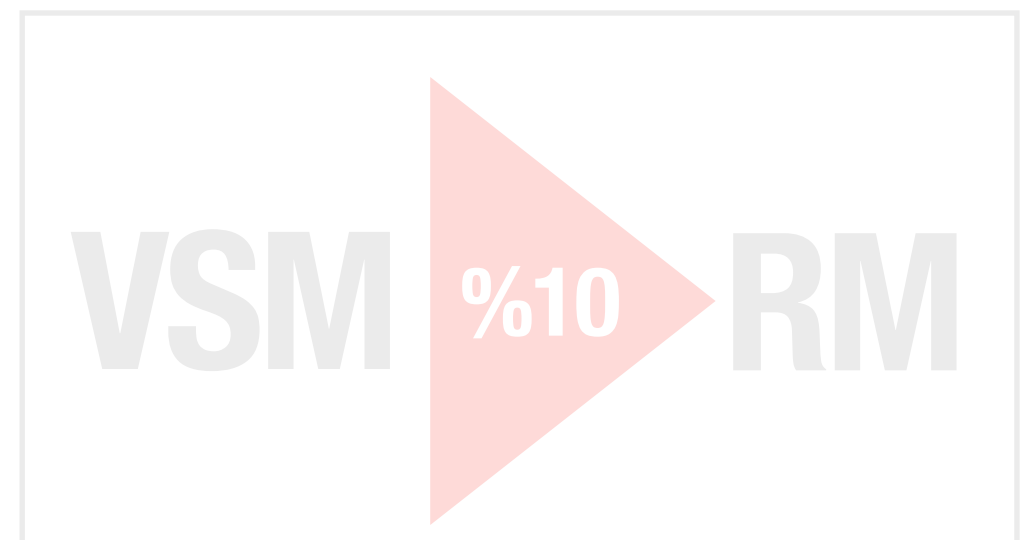




Test Results _ VSM vs RM

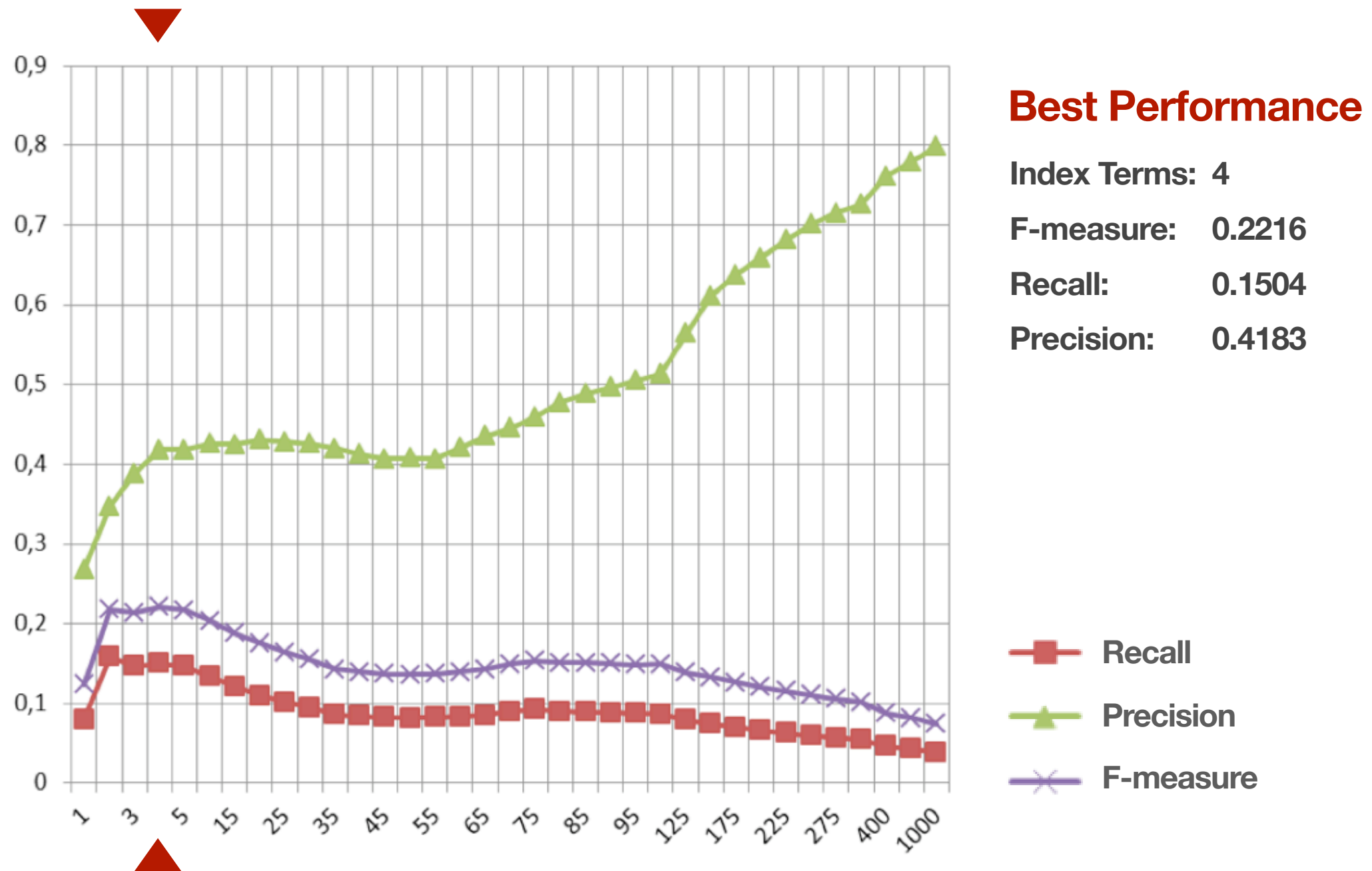
- It appears that the selected best VSM method is more advantageous than the selected best RM method, providing higher recall (%10.17) and precision (%10.77) values

Turkish



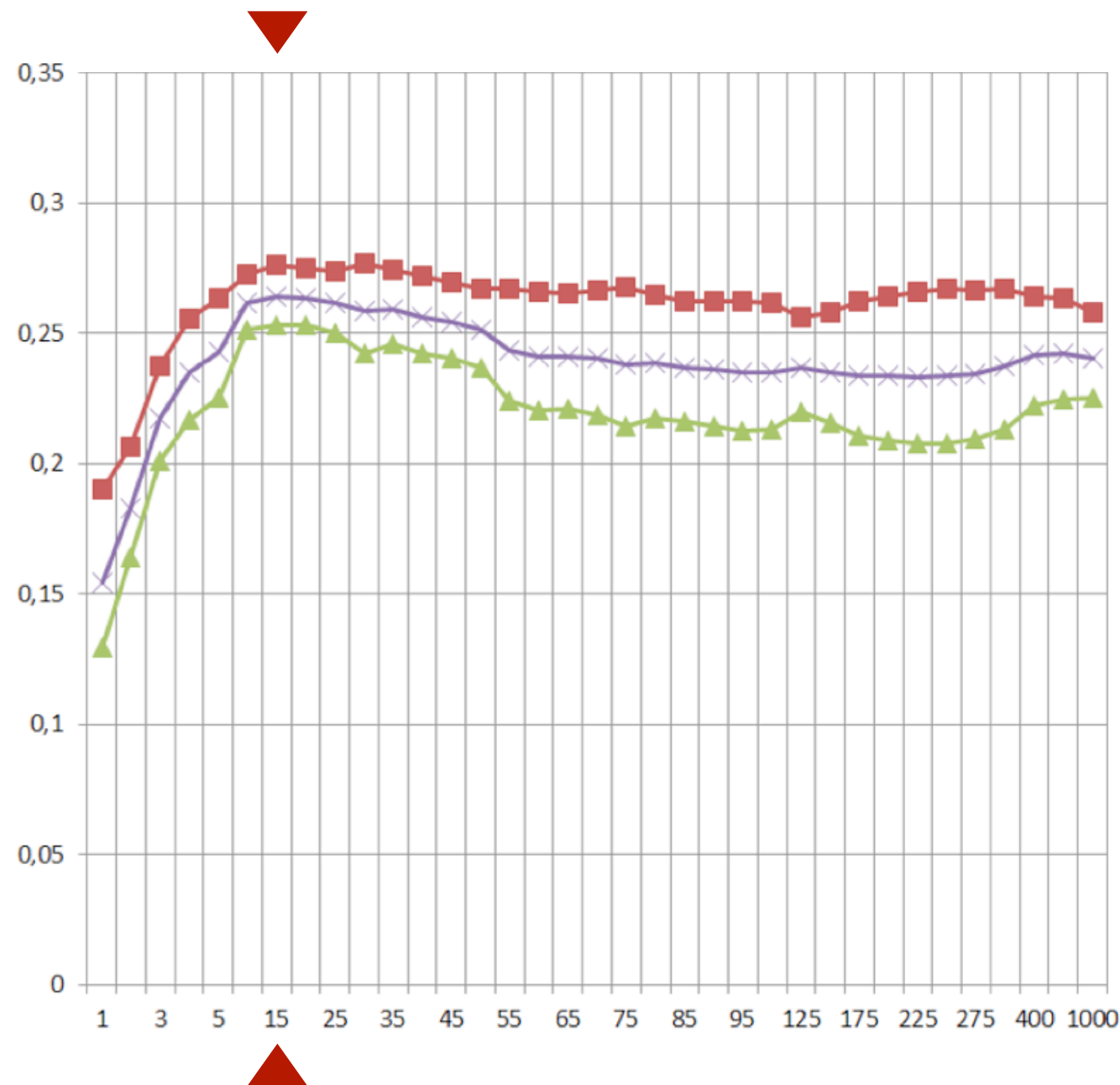


Test Results _ AND combination of VSM and RM





Test Results _ OR combination of VSM and RM





Test Results _ AND vs OR

- ▶ The AND combination of the methods achieved a %7.9 increase compared to the best case with the highest precision value
- ▶ The OR combination of the methods achieved a %1.2 increase compared to the best case with the highest recall value

%7.9

%1.2

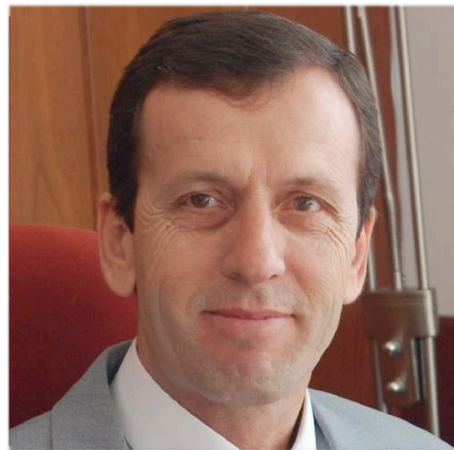


Conclusion

- ▶ SLD that drew special attention in TDT research is applied first time on a Turkish Corpus using two different methods
- ▶ VSM performs better than RM in identifying the similarities of news items on a Turkish Corpus
- ▶ Retrieval performance of SLD algorithms can be increased to some extent by employing both VSM and RM models



Authors



**Yaşar
Tonta**



**Güven
Köse**



**Hamid
Ahmadi**



Thank you for your attention!

Questions?



References

*** please refer to the publication**