Improving Keyword Extraction from Movie Subtitles by Utilizing Temporal Properties

MATÚŠ KOŠÚT

MARIÁN ŠIMKO

Introduction

- Growing importance of automatically annotated content
- Keyword extraction
- Complexity of audio / video processing
- Obtaining subtitles
- Processing and extraction of keywords from subtitles

Current state

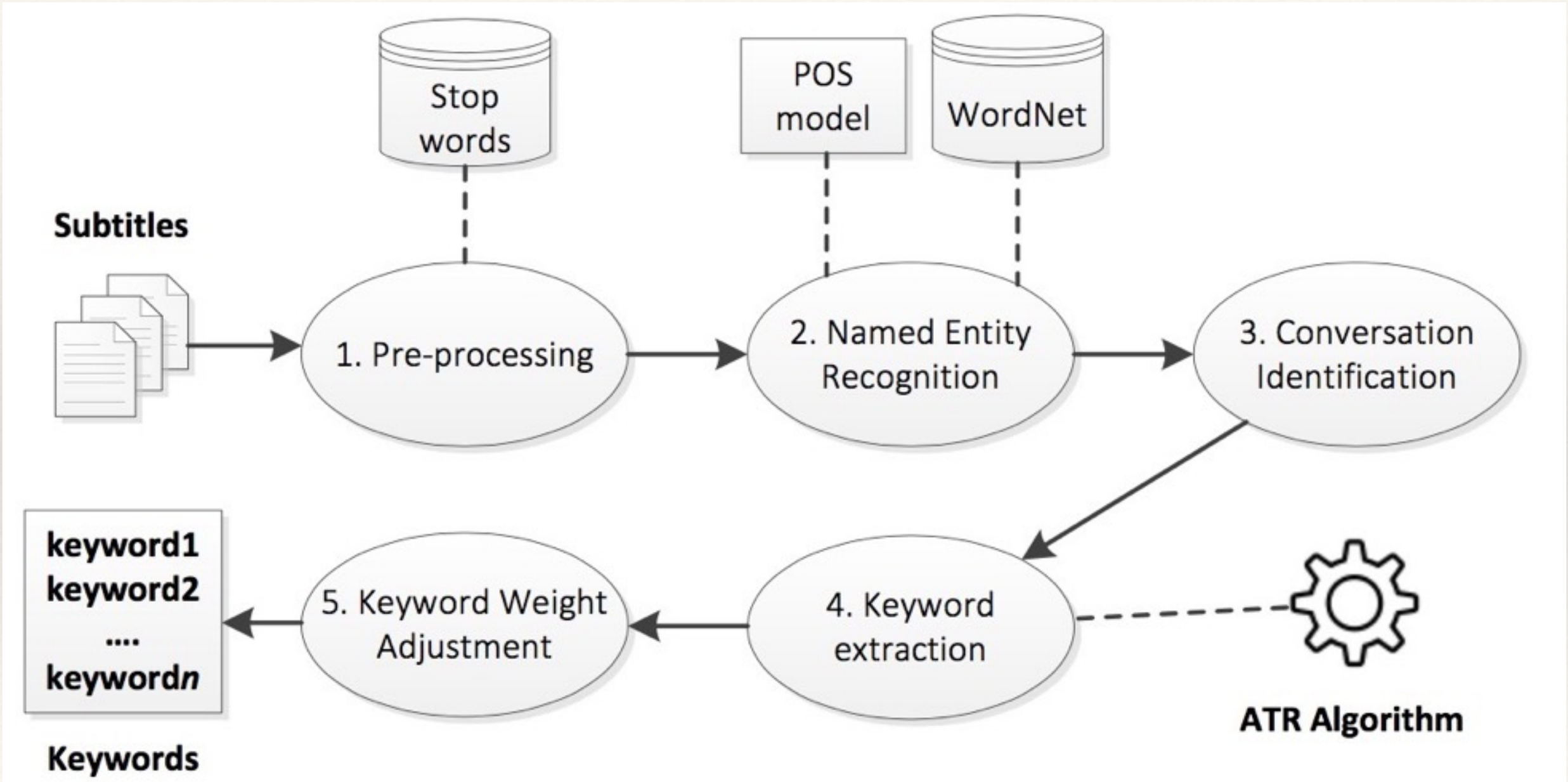
- Keyword extraction from text of subtitles
- Movie classification based on subtitles
- Movie summarisation

- VIRUS (Video Information Retrieval Using Subtitles) (Langlois et al. 2010)
- Video classification based on subtitles (Katsiouli et al. 2007)
- Automatic categorisation and summarisation (Demirtas et al. 2010)

Approach to the solution

- Subtitles as a source of metadata
- Utilizing temporal properties
- Conversation detection
- Named entity recognition

Method

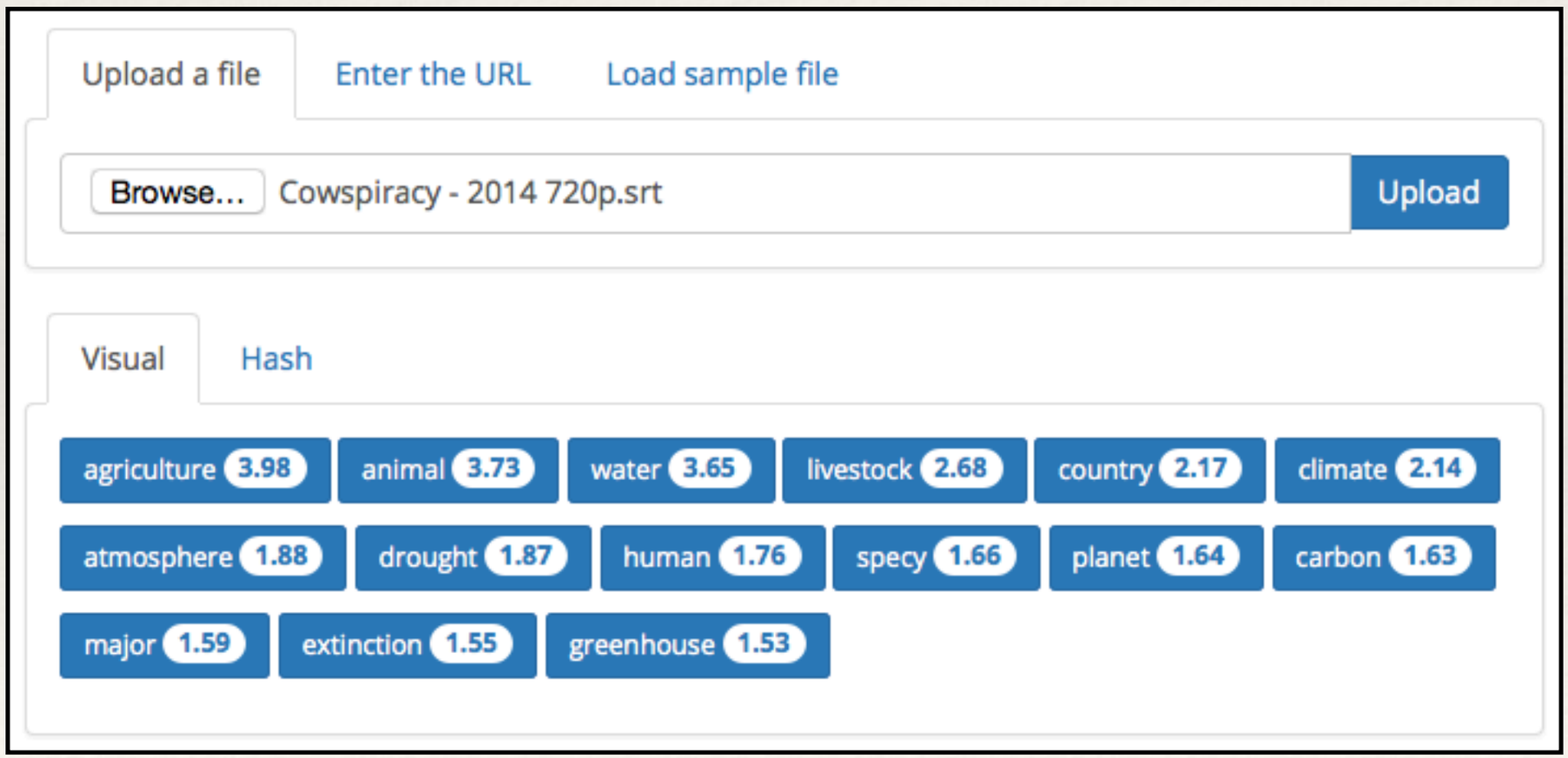

Stop words
POS model
WordNet
Subtitles
1. Pre-processing
2. Named Entity Recognition
3. Conversation Identification
keyword1
keyword2
....
keywordn
5. Keyword Weight Adjustment
4. Keyword extraction
ATR Algorithm
Keywords

Conversations identification

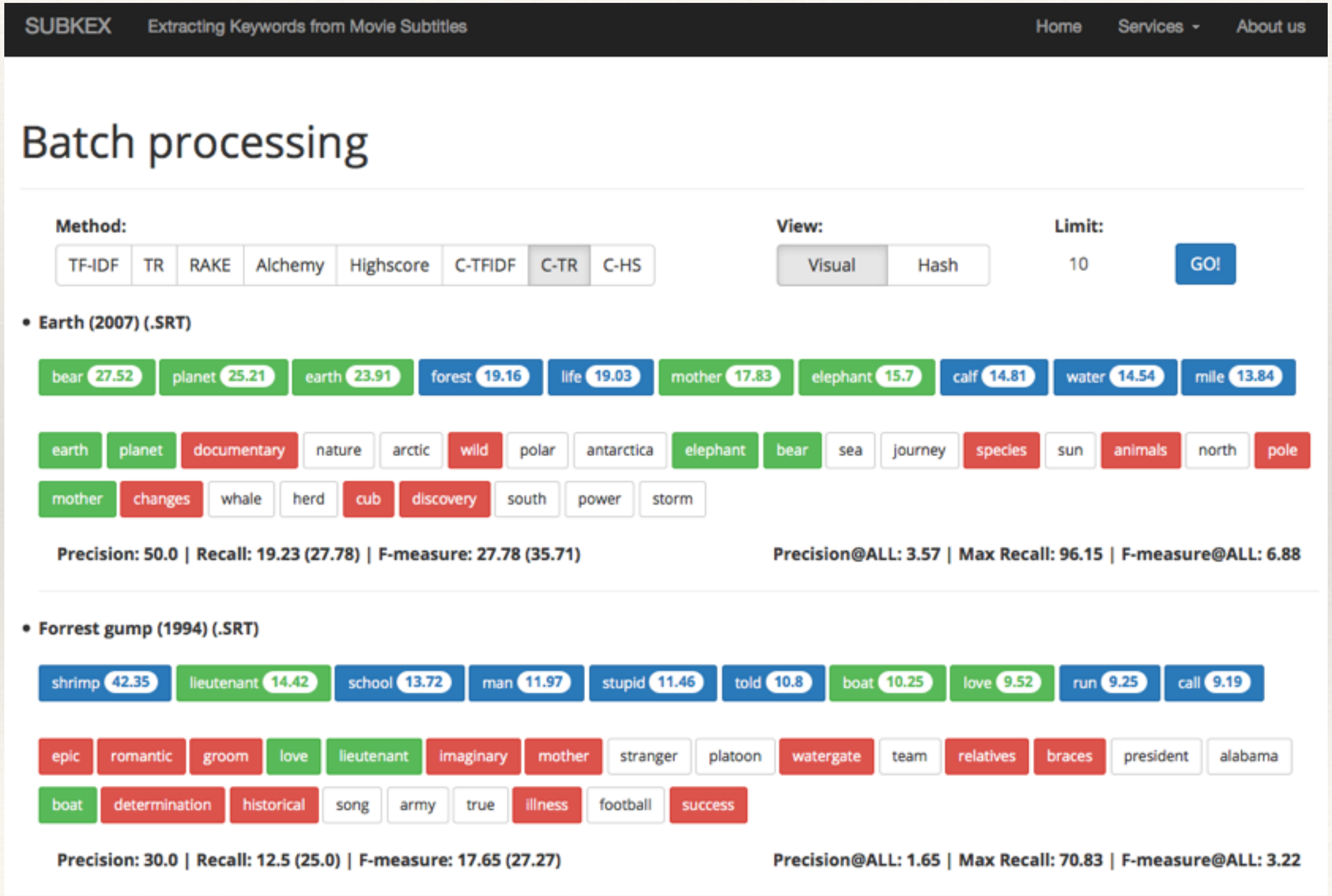
- Dividing movies into blocks based on the time distribution of captions
- Conversation placement in the movie
- Showtime of captions during the conversation
- Normalisation of length according to the speech rate (how much words they say per minute)
- Rating the conversations

Implementation

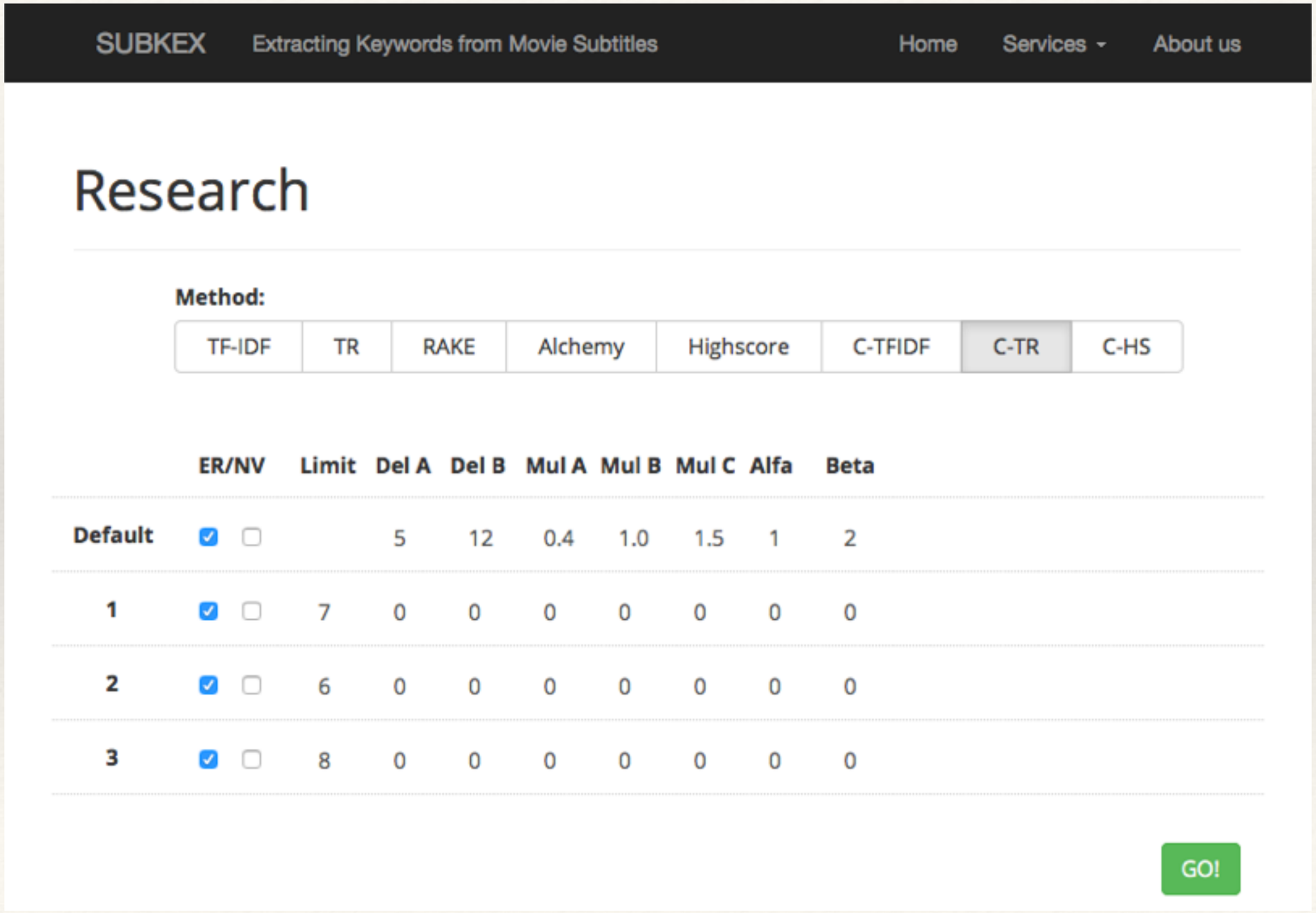
- Web application for keyword extraction



Implementation



Implementation



Evaluation

- We conducted 2 experiments
- Comparing the results of extraction against the golden standard in a priori synthetic experiment
  - Golden standard: keywords acquired from MoviesCus service for movie lookup
  - Gold standard (A priori)
  - User experiment (A posteriori)

Comparison with the Gold Standard

- 200 randomly chosen movies from IMDB database.
- English subtitles acquired from OpenSubtitles community service for sharing the subtitles
- Our methods utilising temporal properties against the basic ATR methods extracting from pure text
- Evaluation of precision, recall, f-score

Comparison with the Gold Standard

Measure @10	P	R	F	R'	F'
TF-IDF	9.75	3.78	5.45	6.05	7.41
TF-IDF+ NER	20.65	8.05	11.58	14.02	16.57
TF-IDF+ NER+C	24.92	9.70	13.96	18.23	20.34
TextRank	19.10	7.41	10.68	12.65	14.80
TextRank+ NER	24.47	9.53	13.72	17.20	19.74
TextRank+ NER+C	**30.30**	**11.80**	**16.99**	21.82	24.60
HighScore	18.04	7.02	10.11	12.99	14.68
HighScore + NER	21.46	8.36	12.03	15.71	17.68
HighScore + NER+C	28.64	11.14	16.04	**23.24**	**24.87**

- TextRank improvement P: 11,20% F: 6,31%
- TF-IDF achieved the best increase P: 15,17% and F 8,51%

User experiment

- 20 newest movies from IDMB top 250
- 20 keywords extracted per movie subtitles
- 17 participants
- Comparing recall of basic ATR algorithms against our methods (altogether)
- Evaluation of precision was based on weighted rating with using 4 degree Likert scale

User experiment

Number of keywords (n)	TF-IDF		TextRank		HighScore	
	Basic	Full	Basic	Full	Basic	Full
1	6.90	69.06	20.69	78.00	10.34	75.06
2	6.90	67.31	17.24	79.00	17.24	68.08
3	9.20	66.13	21.84	76.00	22.99	64.19
4	9.48	64.76	21.55	73.00	20.69	63.21
5	10.34	63.33	19.31	72.00	17.93	62.99
6	13.79	62.22	20.69	69.00	20.11	65.12
7	13.79	61.40	22.17	68.00	20.69	62.52
8	13.36	60.53	21.55	66.00	21.12	61.63
9	13.41	60.22	20.69	67.00	21.46	59.78
10	14.83	58.63	21.38	66.00	21.03	57.99

- Precision at n first results

Conclusions

- Web service for keyword extraction from subtitles
- Traditional ATR algorithms designed to text documents yield poor results with movie subtitles
- Method that is utilising temporal properties of subtitles
  - Conversations detection
  - Named entity recognition
- Both methods succeeded in improving performance of the original ATR algorithms