ATEX: A RULE-BASED SENTIMENT ANALYSIS SYSTEM PROCESSING TEXTS IN VARIOUS TOPICS

Polina Panicheva
EPAM Systems St. Petersburg, Russia
ppolin86@gmail.com

Sentiment analysis

- Started in 1990's in works by J. Wiebe
- Subjectivity, sentiment analysis, polarity classification
- Russia: special in being commerce- and business-oriented
  - Fine-grained tuning capabilities
  - Easy access to rules and output analysis
  - Thus rule-based algorithms appear to fit better

The ATEX sentiment analysis algorithm

- Implemented in EPAM Systems
  - A project in Kazakhstan
  - For Russian news texts
- A rule-based algorithm
- Linguistically rich
  - Processes deep linguistic information
- Participated in the ROMIP 2012 sentiment analysis tracks
  - Results

ATEX: morphological analysis

- A morphological dicationary

  - generated initially with data from [Zaliznyak]

  - contains all word-forms and respective morphological information

- Analysis of unknown words and typos
  - Word-ending: defines normal form and morphology
  - Word-prefix: is deleted and the rest of the word is looked up in the dictionary
  - Fuzzy search in dictionary: Levenstein distance

ATEX: syntactic analysis

- Formal grammar:
  - Homonymy resolution
  - Phrases
  - Bigger phrases, up to complex sentences
- Algorithm applying dependencies to the resulting phrase structure
- Some additional rules applying negation
- Finally: a dependency tree representing a sentence

ATEX: sentiment analysis

- Sentiment: a polarity value for one or more words
  - +1
  - -1
  - 0
  - null
- No sentiment degree: all polarity values are equally “sentimental”

Keyword sentiment

- Keywords
  - *хороший, плохой, неприятный, трус, успех, провал, угроза, позитив, оперативно, своевременно, слишком*

Sentiment rules

- Phrase
  - *пойти навстречу, на лапу, душа компании, по фазе, так себе, промыть мозг, поставить крест, с ума, из ума, ниже плинтуса*
- Inverted sentiment
  - *нет, без, отсутствие, удаление, лишение, отрицание, устранение, отсутствовать, удалять, лишать, отрицать, устранять*
  - Cases of special words
    - Нет проблем
    - Не удается
- Dependency-based rules
  - Degree, loss, problems, lack, flexibility

ATEX results: examples

Sample id	Text	Sentence sentiment
1049	"На данный момент не вижу перспективы(-1) никаких военных действий за исключением мер по защите дипломатических представителей, а также справедливого наказания ответственных за эту ужасную акцию(-1)", - сказал Терци.	-1
1068	"Льоренте все еще принадлежит Атлетику и, похоже, готов играть. Впрочем, мы все равно потеряли(-1) одного отличного(0) футболиста(0) и хорошего(0) человека(0)", - сказал Бьелса, намекая на уход Хави Мартинеса в мюнхенскую "Баварию".	-1
1108	"В период после нашей предыдущей встречи мировая экономика по-прежнему испытывала немалые трудности и продолжает подвергаться рискам падения; финансовые рынки остаются(-1) нестабильными, тогда как высокий уровень(-1) дефицита госсектора и государственной задолженности в некоторых развитых экономиках в значительной мере сдерживает процесс восстановления экономики", - отмечается в документе.	-1

Task description: ROMIP sentiment analysis tracks

- 3-class polarity classification of direct and indirect speech in news texts: positive, negative and neutral (=no polarity)
- 2-class polarity classification of product reviews: positive and negative
- 3-class polarity classification of product reviews: positive, negative and containing significant positive and negative polarity at the same time

Task description: tuning the ATEX system

- No training applied
- Tuning:
  - A corpus of new texts from Russian and Kazakh web resources in Russian: 3000 sentences manually annotated
- Sentence polarity value:
  - Sign of words’ mean value
  - 0 = neutral = null
- 2 modes: with/no sentence boundaries for resulting document polarity

Task description: ROMIP data

Topics	Number of samples	Positive	Negative	Neutral/ containing + and -	Percent of the most frequent class, %
News	4573	1448	1234	1890	41
Movies	408	330	78	-	80
		266	63	79	65
Books	129	112	17	-	87
		100	9	20	78
Cameras	411	397	14	-	97
		351	7	53	85

ATEX results: evaluation

- Evaluation: F-measure, Accuracy, Precision, Recall
- Unbalanced data
- -> F-measure appears adequate
- Example:
  - 90% documents negative,
  - A goldstandard approach assigning the most frequent class achieves 90% accuracy
  - Although totally insensitive to the other class
  - While F-measure accounts for Recall and Precision for all classes

ATEX results: news samples

Number	System ID	Object	Classes	Precision Macro	Recall Macro	F_Measure Macro	Accuracy	
1	xxx-4	news	3	0.626	0.616	0.621	0.616	
2	**ATEX**	**news**	**3**	**0.606**	**0.579**	**0.592**	**0.571**	**with no sentences**
3	**ATEX**	**news**	**3**	**0.606**	**0.576**	**0.590**	**0.569**	**with sentences**
4	xxx-5	news	3	0.579	0.568	0.574	0.575	

ATEX results: blog reviews

- Movies:
  - 2-class: 1st , 0.707 F-measure. Rel. lower accuracy – 0.806
  - 3-class: 2nd, 0.503 F-measure. Relatively lower accuracy – 0.596
- Cameras:
  - 2-class: 5th,
  - 3-class: 5th,
- Books:
  - 2-class: 3rd,
  - 3-class: 6th.

Conclusions

- Relatively high results in native news sample
  - Despite difference in sources and time in tuning/testing data
- Considerable results in all other topics
  - Highest performance in movie reviews
  - More modest in books, cameras.
- -> the results reflect the degree of technical information relevant to sentiment in the topic:
  - Movies and news appear the most common and the closest to each other
  - Books and especially cameras include more special terms in their sentiment expressions
- The rule-based algorithm with no training performs with high replicability
  - High F-measure
  - High results in various topics

Future work

- Statistical account of rules applied:
  - Retrieve most significant rules
  - Retrieve important differences between topics
  - A step towards automatic lexical rule filling
- Model a neutral class and a class containing both polarity values
  - ML appears to draw fine lines between plarity, neutral and containing both

Thank you!

- Questions?