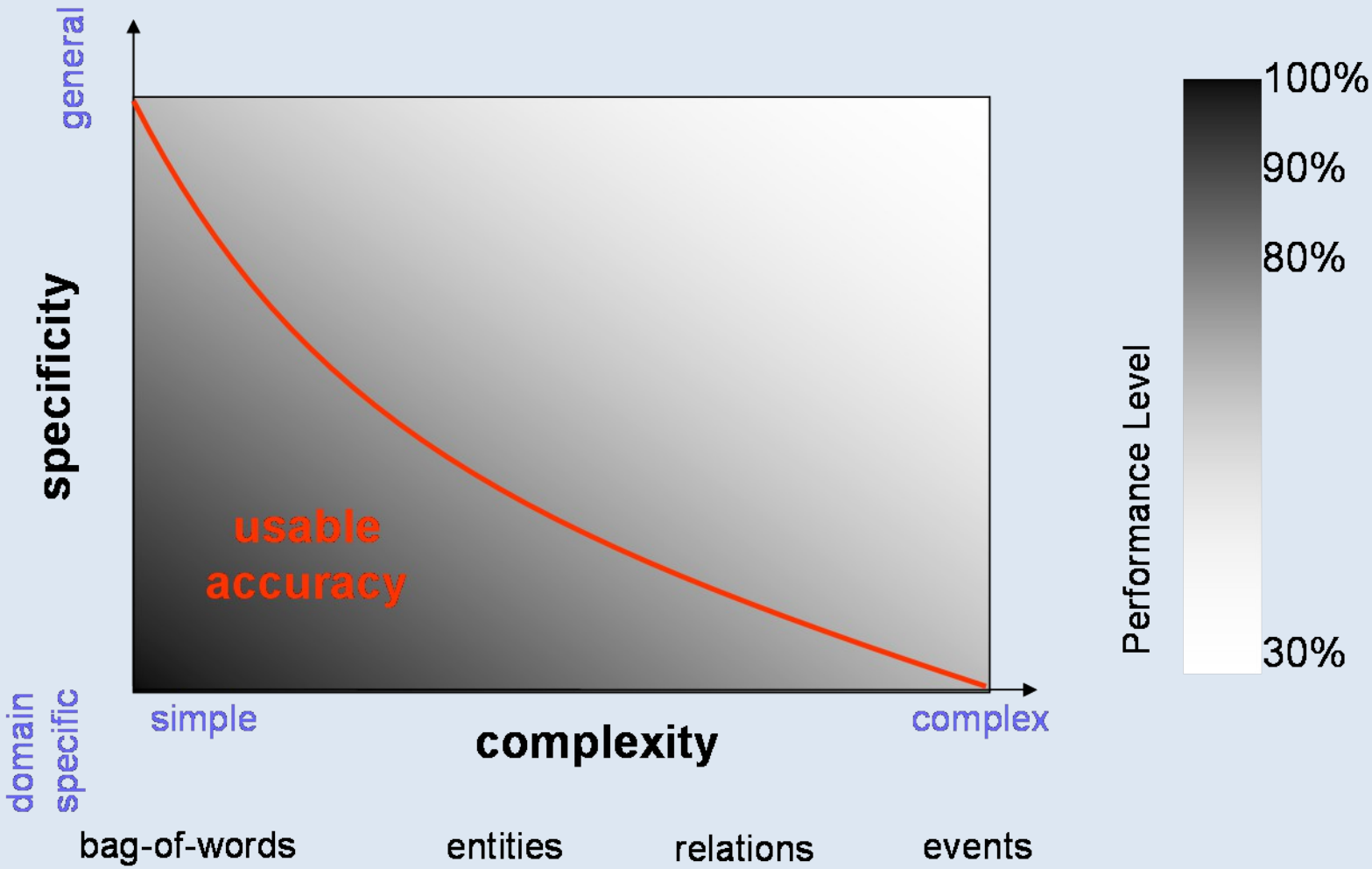
A tailor-made one-size-fits-all approach to sentiment analysis

Diana Maynard
University of Sheffield, UK

What's the problem?

- Like most language engineering tasks, sentiment analysis works best on a specific domain, in a single language, where the task is clear and the range of options limited
- This means that tailor-made approaches to SA are the key to success
- Off-the-shelf tools for sentiment analysis are unlikely to work well when used for different tasks
- This is great when you want to find opinions about the latest films from a single film review site, written in English, as you can train models for this purpose and get reasonable results
- But what if you want to do more complex forms of analysis on a potentially unknown set of documents containing potentially unknown kinds of entities?

Language processing tradeoffs



The task

- Two very different domains:
  - Greek and Austrian Parliamentary texts
  - German rock concerts
- We want to find out ultimately:
  - What are the opinions on crucial social events and who are the key people involved?
  - How are these opinions distributed in relation to demographic user data?
  - How have these opinions evolved?
  - Who are the opinion leaders?
  - What is their impact and influence?
- The first task is to investigate public opinion about key entities and events

Sentiment analysis is hard...

- And it's even harderin our case because:
  - we have lots of different text types and domains
  - we're processing social media
  - we're processing multiple languages
  - we don't necessarily know what we're looking for
- Experimenting with different techniques and subcomponents to build up a complex system
- Experimenting with techniques for cross-language adaptation

GATE to the rescue...

- We use GATE, a tool for language engineering, because it is robust and adaptable with a component-based architecture
- Lots of experience adapting IE applications to different languages and domains
- Can we do the same thing for sentiment analysis?
- Main challenges:
  - how can we easily port sentiment analysis to a new language, and how can we process documents in multiple languages most efficiently?
  - how can we easily adapt applications which work on standard reviews to deal with noisy social media?

We use a rule-based approach because...

- Although ML applications are typically used for Opinion Mining, this task involves documents from many different text types, genres, languages and domains
- This is problematic for ML because it requires many applications trained on the different datasets, and methods to deal with acquisition of training material
- Aim of using a rule-based system is that the bulk of it can be used across different kinds of texts, with only the pre-processing and some sentiment dictionaries which are domain and language-specific

Application Components

- Structural pre-processing, specific to social media types
- Linguistic pre-processing (including language detection)
- Standard language-specific NE, term and event recognition
- Additional targeted language- and task-specific gazetteer lookup
- JAPE grammars for opinion finding
- Aggregation and summarisation of opinions

Conditional processing in GATE

- In GATE, you can set a processing resource in your application to run or not depending on certain circumstances
- You can have several different PRs loaded, and let the system automatically choose which one to run, for each document.
- This is very helpful when you have texts in multiple languages, or of different types, which might require different kinds of processing
- For example, if you have a mixture of German and English documents in your corpus, you might have some PRs which are language-dependent and some which are not
- You can set up the application to run the relevant PRs on the right documents automatically.

Conditional processing with different languages

- Suppose we have a corpus with documents in German and English, and we only want to process the English texts.
- First we must distinguish between the two kinds of text, using a language identification tool
- We use TextCat (a GATE plugin) to add a feature on each document, telling us which language the document is in
- Then we run a conditional processing pipeline, that only runs the subsequent PRs if the value of the language feature on the document is English
- The other documents will not be processed
- What if we want to process both German and English documents?
  - we can just call some language-specific PRs conditionally, and use the language-neutral PRs on all documents

What if the documents themselves are in multiple languages?

- Segment Processing PR enables you to process labelled sections of a document independently, one at a time
- It then merges back the individual sections once they've been processed
- Useful for
  - when you want annotations in different sections to be independent of each other
  - when you only want to process certain sections within a document
  - processing multiple languages separately within a single document

English and German content

Annotation Sets | Annotations List | Annotations Stack | Co-reference Editor | Text

15.03.2011
10:43 Uhr
The Toilets are called "ToiToi" and they were very simple porta Toilets.
But ther are some Clean stations, in which you can take a shower and you can go to toilet, but this station cost money.

I dont know yet how much the shower cost.
Ah and you shower allone
And you can crap also in de forest ,if you have paper There is a nice feeliing

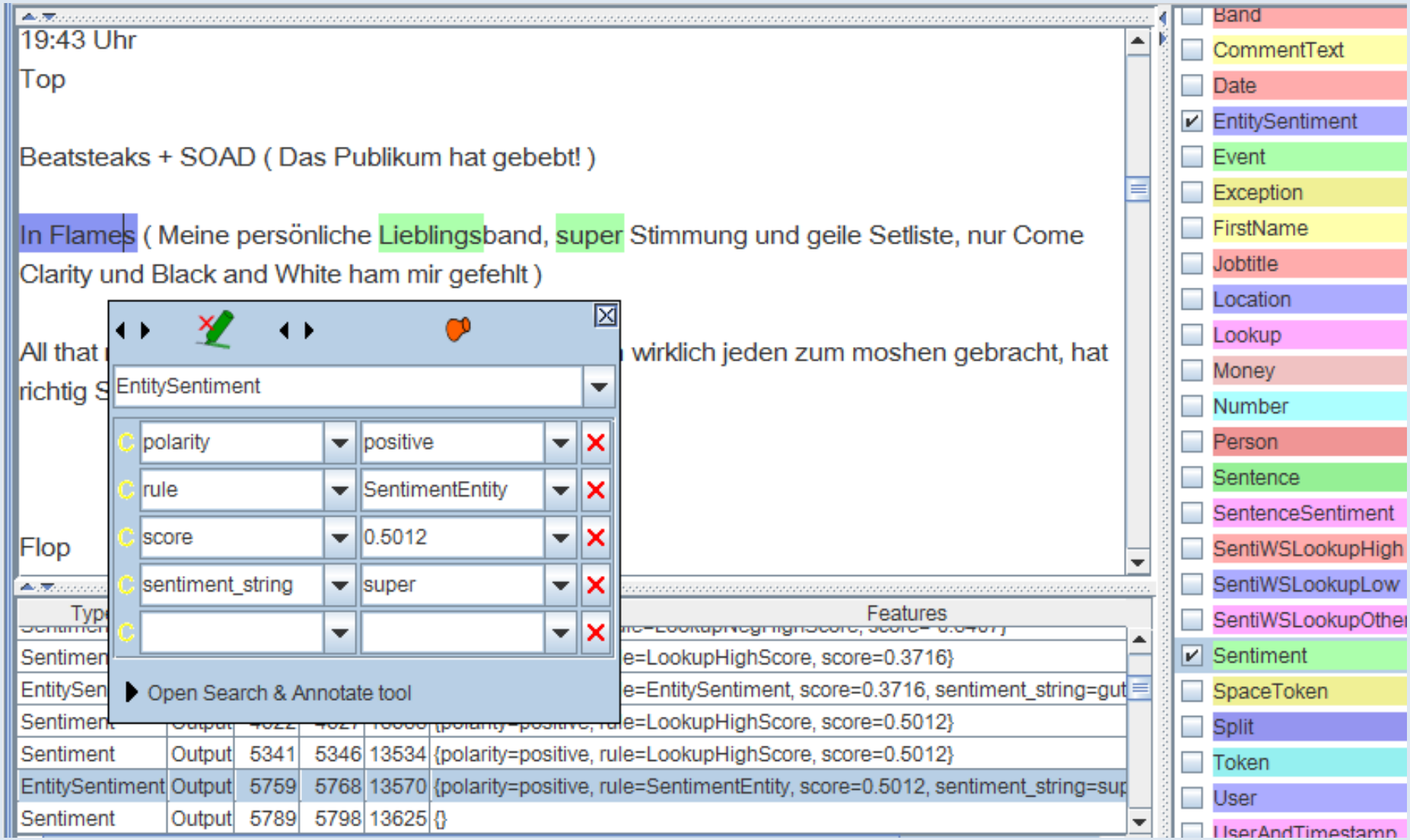
EDITH: Da war jemand schneller, und ein paar miserable Fehler verbessert

One hint: Take Toilet paper with you

Type	Set	Start	End	Id	Featu
EnglishContent	Opinions	1297	1369	53785	{rule=GetEnglishContent}
EnglishContent	Opinions	1370	1489	53786	{rule=GetEnglishContent}
EnglishContent	Opinions	1491	1532	53787	{rule=GetEnglishContent}
EnglishContent	Opinions	1533	1557	53788	{rule=GetEnglishContent}
EnglishContent	Opinions	1558	1636	53789	{rule=GetEnglishContent}

- [ ] Conjunction
- [ ] ConsequentIndicator
- [ ] EmbeddedHead1
- [ ] EmbeddedHead2
- [x] EnglishContent
- [ ] Entity
- [ ] EntitySentiment
- [ ] FirstPerson
- [x] GermanContent
- [ ] Location
- [ ] Lookup
- [ ] Not
- [ ] Number
- [ ] Organization
- [ ] PRP
- [ ] Participle
- [ ] Person
- [ ] PhraseBreak
- [ ] Possessee
- [ ] Possessor
- [ ] Preposition

Sentiment Finding in Rock am Ring



Creating language-specific resources

- Linguistic pre-processing
  - standard tools (POS tagging, language ID, etc)
  - NE and term recognition: existing language-specific plugins
- Gazetteers
  - Automatic translation of gazetteer lists
  - Collection of lists from the web, e.g. swear words, typical phrases
  - Automatic gazetteer induction from texts: bootstrapping approach
- Everything else is language-independent

Creating domain-specific resources

- Unlike the language issue, documents are only ever of a single type
- Social media poses challenges for linguistic processing
  - requires specially developed PRs with more flexible matching, special handling of tweets etc (tokeniser, POS tagger etc)

  "RT @Bthompson WRITEZ: @libbyabrego honored?! Everybody knows the libster is nice with it...lol...(thankkkks a bunch;))"

  - See our work on the TrendMiner project for more details of this problem and how we handle it
  - Phenomena like sarcasm occur more often on less formal texts
- We can again use conditional processing on a document-by-document basis to deal with these issues

Conclusions

- It's best if you can tailor your application as much as possible ot the domain and language
- But if you have to process multiple kinds of text, a modular rule-based approach can allow you to combine the specific resources with the generic resources
- GATE is ideally set up for this (of course, other tools are available too...)
- We use a rule-based approach, but lots of current research on automatic induction of new training data on different kinds of text

More information

- GATE http://gate.ac.uk (general info, download, tutorials, demos, references etc)
- The EU-funded ARCOMEM and TrendMiner projects are dealing with lots of issues about opinion and trend mining from social media, and use GATE for this.
  - http://www.arcomem.eu
  - http://www.trendminer-project.eu/
- More information on dealing with the problems of social media:
  - D. Maynard and K. Bontcheva and D. Rout. Challenges in developing opinion mining tools for social media. In Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012, May 2012, Istanbul, Turkey.


Personalisation?
Will this do for all of you?
No - that won't fit ME
1 SIZE FITS ALL
We are ALL unique with individual needs and requirements
© ogilviedesign.co.uk