Automatic identification of general and specific sentences by leveraging discourse annotations

-Annie Louis and Ani Nenkova

Presentation by Vinay Krupakaran
Feb 24, 2015

An Example

"George RR Martin is the worst wedding planner in history!"

An Example (contd.)

“G.R.R. Martin, the author of the series, ‘A song of Ice and Fire’ has been touted as the worst wedding planner in history due to his penchant for killing off important characters during the ceremonies.”

A more serious example..

The Booker prize has, in its 26-year history, always provoked controversy.

The novel, a story of Scottish low-life narrated largely in Glaswegian dialect, is unlikely to prove a popular choice with booksellers who have damned all six books shortlisted for the prize as boring, elitist and - worst of all - unsaleable.

Potential uses

- Prediction of writing quality.

Prescriptive books on writing advise to avoid sentences that use vague or abstract words or follow them up with specific clarifications

- Text generation systems.

Control the type of content produced

- Information extraction systems.

Distinguish between types of information

Previous Work

- Reiter and Frank (2010): An automatic approach to distinguish between noun phrases which describe a class of individuals (generic) versus those which refer to a specific individual(s).
  a. The lion chased the deer down.
  b. Lions can eat up to 30 kg. in one sitting.

- Mathew and Katz (2009): Distinguish sentences which relate to a specific event (called episodic) from those which describe a general fact (habitual sentences).
  a. Italians drink wine.
  b. The Italian hosted the dinner.

Aim

- To design a supervised classifier to distinguish between General and Specific sentences
- Evaluate the classifier by checking its performance on a specific task.

Training data

- No existing corpus for specificity
- Use certain discourse relations annotated in the Penn Discourse Tree Bank (PDTB):
  a. Specification
  b. Instantiation

Training data (contd.)

- Specification: || Arg 1 || <— || Arg 2 ||

Argument 2 restates Arg1- describes it in more detail

Eg. “This is an old story. *In fact we’re talking about years ago before anyone heard of asbestos having any questionable properties*.”

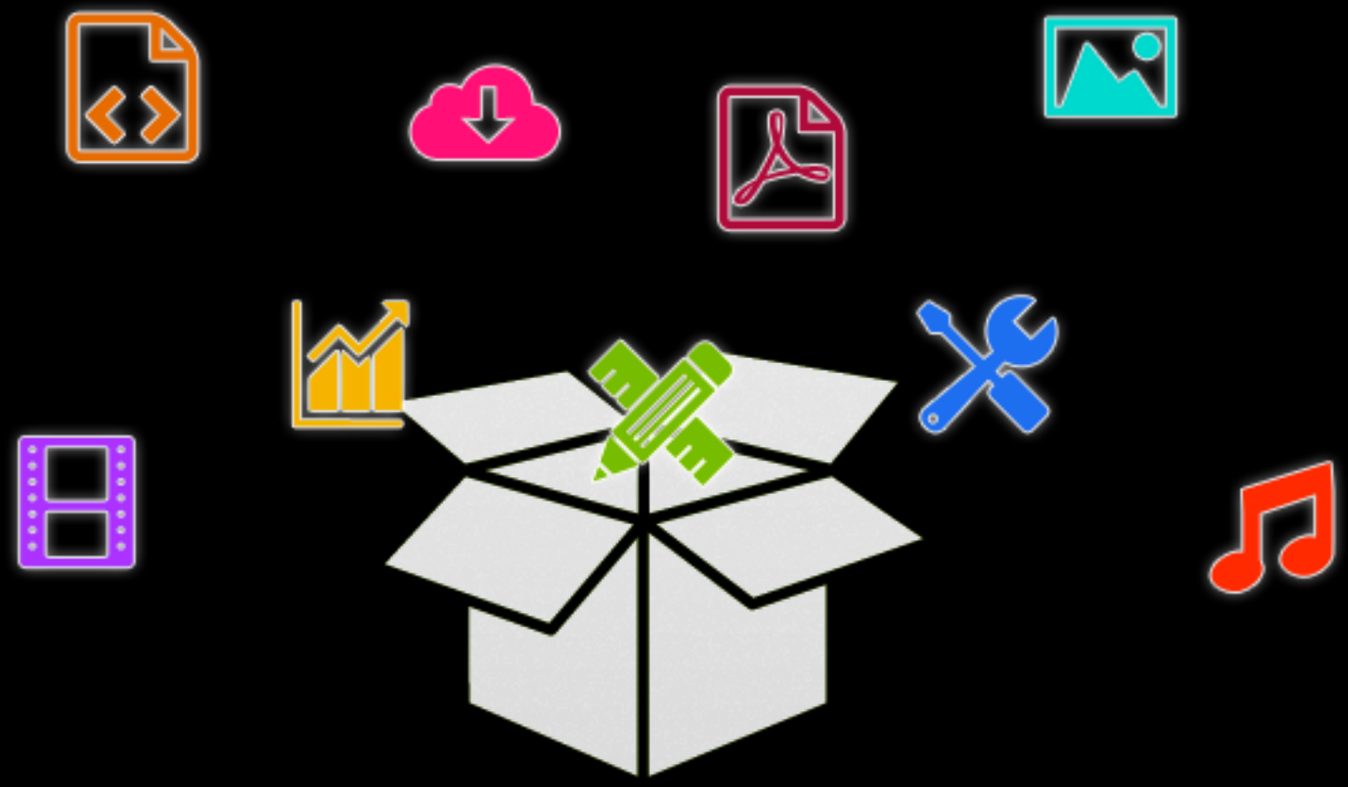
- Instantiation: Argument 1 evokes a set. Argument 2 describes it in further detail : || Arg 1 || ^ || Arg 2 || must hold.

Eg. He says he spent $300 million on his art business this year. *In particular, a week ago, his gallery racked up a $23 million tab at a Sotheby’s auction in New York buying seven works, including a Picasso*.

Training data (contd.)

- We do not care about the relationship between the pairs of these sentences
- There are qualities that make a sentence *'general'* regardless of what follows it.
- First sentence as the general sentence
- Second sentence as the specific sentence

Features!



Set of features

- Sentence length - General sentences are shorter
- Polarity - Sentences with strong opinions are typically general.
- Specificity - Specific sentences = more specific words.
  a. WordNet : Hypernym relations.
  b. IDF : Inverse document frequency for a word. [log N/n]
- Numbers and Symbols - Specific sentences contain more numbers and symbols like the pound or the dollar.

More Features!

- Language Model - General sentences contain more unexpected phrases. Eg. Despite recent declines in yields, investors continue to *pour cash* into money funds. Log probability and perplexity of the sentences used.

- Syntax - Specific sentences have longer verb phrases. Number and length of verb phrases. General sentences have more qualitative words like adjectives and adverbs.

- Words - Count of each word in a sentence. Words not seen in the training set are ignored.

Results

Features	Instantiations	Specifications
NE+CD	68.6	56.1
language models	65.8	55.7
specificity	63.6	57.2
syntax	63.3	57.3
polarity	63.0	53.4
sentence length	54.0	57.2
all non-lexical	75.0	62.0
lexical (words)	74.8	59.1
all features	75.9	59.5

Table 1: List of features used and their individual performances

- Two classifiers - Trained on Instantiations and Specifications
- Use a Logistic Regression - Probabilities are more appropriate than hard classification

- Instantiation classifier outperforms the Specifications classifier
- Best feature - Combination of all features
- Best class of features - Words

Feature Analysis

Top word features that appear in at least 25 training examples :

- General : number, but, also, however, officials, some, what, prices, made, lot, business, were
- Specific : one, a, to, co, i, called, we, could, get, and, first, inc
- Domain specific words : officials, prices, business, number / co, inc

Assumptions made for the features were mostly true.

- Numbers and names present in specific sentences
- Plurals were present in general sentences and so on.

Unexpected outcome: The dollar sign is more frequent in the general sentences.

Testing on new sentences

- Three articles from the WSJ - PDTB
- Six from the Associated Press - AQUAINT corpus
- Two from the Financial Times - AQUAINT corpus
- Manually annotated using Amazon's MTURK (crowdsourcing internet marketplace)
- Judgements from five unique users for each sentence - General/Specific/ Can't decide

	WSJ articles			AP articles		
Agree	**total**	**gen**	**spec**	**total**	**gen**	**spec**
5	96	51	45	108	33	75
4	102	57	45	91	35	56
3	95	52	43	88	49	39
undecided	1			5		
Total	294	160	133	292	117	170

Table 2: Annotator agreement

Results

		WSJ sentences				AP sentences		
Examples	**Size**	**All features**	**Nonlexical**	**Words**	**Size**	**All features**	**Nonlexical**	**Words**
Agreement 5	96	90.6	96.8	84.3	108	69.4	94.4	78.7
Agreement 4 + 5	198	80.8	88.8	77.7	199	65.8	89.9	74.8
Agreement 3 + 4 + 5	293	73.7	76.7	71.6	287	59.2	81.1	67.5

Table 3 : Results on the annotated data

- Non - lexical features give the best individual performance
- Contrary to the previous results, lexical features now give a lower accuracy

- Lexical features cannot account for every example type
- Non - lexical features provide better coverage and portability across corpora
- Results better on sentences with higher agreement

Task based evaluation

- Manually created summaries for a set of news articles.
- Data obtained from the Document Understanding Conference (DUC) in 2005.
- A topic statement and 25-50 news articles on the topic were used to create gold standard summaries of the topic
- The annotators specified, for each input, what kind of summary would be appropriate.
- These texts were then given to trained assessors who wrote general or specific summaries accordingly.
- 146 general and 154 specific summaries

Performance on task based evaluation

Text	General category	Specific category
Summaries	0.55 (0.15)	0.63 (0.14)
Inputs	0.63 (0.06)	0.65 (0.04)

Table 4: Mean and SD values of specificity levels for inputs and summaries

T - test results:

- Summaries : Difference in values has a p-value of1.5e-06
- Inputs : Difference in values has a p-value of 0.275

- The system can distinguish between general and specific summaries

Conclusion

- Introduction of a new task : General vs Specific sentences
- How discourse relations can be used to obtain data for our problem
- Introduction of features that can be used to evaluate the specificity of a sentence with high classification performance results over the baseline
- Created a classifier that can distinguish between general and specific summaries written by people.


?
Q
&
&
A
!