Institut Mines-Télécom

Uncertainty detection in natural language

Pierre-Antoine Jean (LGI2P)
Sébastien Harispe (LGI2P)
Sylvie Ranwez (LGI2P)
Patrice Bellot (LSIS)
Jacky Montmain (LGI2P)

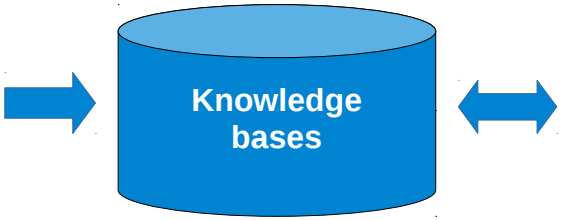
Positioning

General context

Take into account uncertainty in enrichment and utilization of knowledge bases for decision making.

Enrichment

- **Information extraction**
  - + Named entity recognition
  - + Relation extraction
- **Uncertainty detection**
  - + Binary detection
  - + Scope of the uncertainty



Utilization

- **Knowledge inference**
  - + Knowledge discovery
  - + Rules mining
  - + Inconsistency detection
- **Decision making**
- **Question answering**

Uncertainty markers

Uncertainty *

Semantic uncertainty

Semantically uncertain propositions cannot be assigned a truth value *i.e.* it cannot be stated for sure whether they are true or false, given the speaker's current mental state

« *It <u>may</u> be raining.* »

Discours-level uncertainty

Lack intentionnal or not of information from the speaker (source, imprecision, subjectivity)

« *<u>It has been suggested</u>* [***by whom?***] *that he should have involved Clinton much more heavily in his campaign.* »

- **Markers:**
  - **Auxiliary** (verbes modaux): *may, should, would, might*
  - **Speculative verbs** : *suspect, seem, presume*
  - **Adjectives and adverbs** : *probably, likely, perhaps, often*
  - **Fuzzy numeric expressions** : *many, most, some, Experts say, some people*

* Szarvas, G., Vincze, V., Farkas, R., Móra, G., & Gurevych, I. **Cross-genre and cross-domain detection of semantic uncertainty**. Computational Linguistics. 2012.

Results CoNLL 2010 – *Task* 1

	BioScope [1]	WikiWeasel [2]
Uncertainty dimension	Semantic	Semantic / Discourse-level
Best F-measure de CoNLL 2010	Tang et al. – **86,4 %** (55 % on WikiWeasel)	Georgescul et al. – **60,2 %** (78,5 % on BioScope)

Summary

- The kind of texts
- The dimension of uncertainty

} impact methods efficiency

[1] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik: **The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts.** BMC Bioinformatics, 9(Suppl 11):S9. 2008.

[2] Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. **The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text**. Conference on Computational Natural Language Learning (p. 1-12). 2010.

Overview

Training

Labelled sentence n°1
Labelled sentence n°2
...

Feature engineering

	Feature n°1	Feature n°2	Feature n°3	...
Sentence n°1	0,32	0,5	0,02	...
Sentence n°2	0,1	0,12	0,01	...

Support Vector Machine

SVM Model

Prediction

Sentence n°1
Sentence n°2
...

Support Vector Machine

Sentence n°1 – Uncertain
Sentence n°2 – Certain
...

Features

Overall feature				
- Length of the sentence				
Local features				
	Type	Length	Context	Aggregation
Feature 1	Stem	1	Uncertainty markers	Sum
Feature 2	Stem	2	Uncertainty markers	Sum
Feature 3	Stem	1	In uncertain sentence	Sum
Feature 4	PoS	5	In uncertain sentence	Sum
Feature 5	Stem	1	In uncertain sentence	Max

Feature 1 : TCF-1 might play a role in the establishment of ...

→ Sum w1 + w2 ...

Feature 2 : TCF-1 might play a role in the establishment of ...

→ Sum w1 + w2 ...

Feature 4 : NN MD VB DT NN IN DT NN IN ...

→ Sum w1 + w2 ...

Probabilist measure

	Type	Lenght	Context	Aggregation
Feature 1	Stem	1	Uncertainty markers	Sum
Feature 3	Stem	1	In uncertain sentence	Sum

Conditional probability

$S_{i,train} = w_1 \; w_2 \; w_3 \; \ldots \; w_n$

	$\#_{context}$	$\#_{tot}$	%
w_1	1	4	25
w_2	1	1	100
w_3	100	100	100

Feature 1

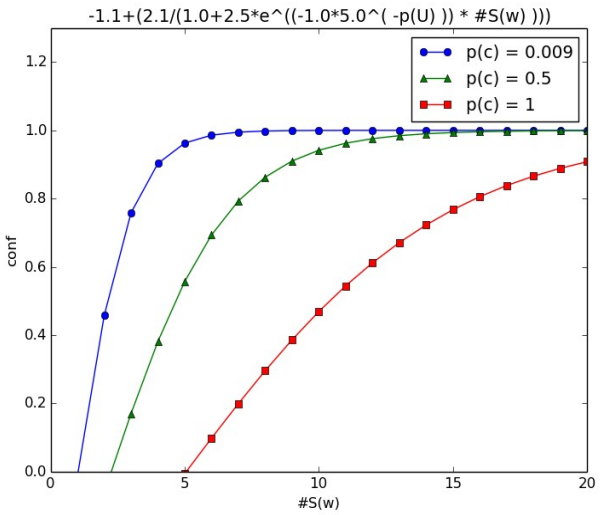
$S_{j,eval} = w_1 \; w_2 \; w_3 \; \ldots \; w_n$

- Uncertainty marker unigrams F_1

$$F_1 = \sum_{k=1}^{n} p\left(w_k \in I_{Su}\right) \times conf_{I_{Su}}\left(w\right)$$

Confidence score

- Allows to evaluate the relevance of specific ngram in taking into account its **number of occurrences** in the corpus and the **probability** to pick randomly an uncertainty marker in the corpus.

Binomial law	Sigmoid function	Control measure
$p(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$		$conf(w) = 1 - \frac{1}{S(w)}$

Results with the conditional probability

	BioScope (86,4 %)	WikiWeasel (60,2 %)	SFU corpus (92,3 %)
Binomial law	79,4	61,4	92,2
Sigmoid	**79,5**	61,7	**92,3**
1 – 1 / s(w)	**79,5**	**62,8**	**92,3**
Without confidence	79,4	59,2	**92,3**

Tang et al. BioScope : 86,4 %
WikiWeasel : 55 %
} Average : 70,7 %

Georgescul et al. BioScope : 78,5 %
WikiWeasel : 60,2 %
} Average : 69,35 %

Notre méthode BioScope : 79,5 %
WikiWeasel : 62,8 %
} **Average : 71,2 %**

- Large disparity between corpuses (nature, number and distribution of uncertainty markers)
- Wikiweasel data are more sparse
- The p probability (success probability) is not required to improve results

Results with many metrics

	BioScope	WikiWeasel	SFU
Conditional probability, 1 – 1 / s(w)	79,5 %	62,8 %	92,3 %

Measures	Confidence	BioScope	WikiWeasel	SFU
PMI	log(#s(w))	75,6 %	33,8 %	88,3 %
	1-1/#s(w)	77,3 %	40,6 %	91,1 %
	Binomial law	76,6 %	52,3 %	91,5 %
	Sigmoid	77,1 %	37,7 %	91 %
	Without confidence	76,4 %	35,1 %	90,6 %
Odds Ratio	log(#s(w))	78,1 %	45,5 %	91,1 %
	1-1/#s(w)	**79,3 %**	52 %	92,1 %
	Binomial law	**79,3 %**	**55 %**	**92,2 %**
	Sigmoid	**79,3 %**	51,5 %	92,1 %
	Without confidence	79,2 %	51,3 %	92,1 %
CPD	log(#s(w))	70,8 %	45,2 %	78,6 %
	1-1/#s(w)	70,4 %	49,9 %	78 %
	Binomial law	69,7 %	48,1 %	80,1 %
	Sigmoid	70,5 %	48,6 %	78,1 %
	Without confidence	69,6 %	48 %	73,3 %
Wllr	log(#s(w))	53,7 %	16,5 %	69,8 %
	1-1/#s(w)	55,1 %	11 %	66,3 %
	Binomial law	55,5 %	45 %	67,1 %
	Sigmoid	55,1 %	11,6 %	65,8 %
	Without confidence	55,1 %	18,9 %	65,7 %

Conclusion

- Short vectorial representation of sentences realized by at most 6 features becoming to a specific aggregation of n-gram's weights
- Weighted sentence's n-grams with their conditional probability to belong to a specific class *c* (according to their context)
- Modelisation of a confidence score to distinguish more relevant n-grams
- Improvement of F-measure on WikiWeasel and the average of F-mesures on the corpus used during CoNLL 2010

Thank you

- Vectorial representation of sentence
- Weighted sentence's n-grams
- Confidence score to distinguish n-grams
- Improvement of F-measure of CoNLL 2010

Pierre-Antoine Jean – pierre-antoine.jean@mines-ales.fr

Sébastien Harispe – sebastien.harispe@mines-ales.fr

Sylvie Ranwez – sylvie.ranwez@mines-ales.fr

Patrice Bellot – patrice.bellot@lsis.org

Jacky Montmain – jacky.montmain@mines-ales.fr