Softwareproject Topics

Katja Markert: Topics SS22

Computerlinguistik
Universität Heidelberg

Topic Suggestions

Variations

Most topics can be handled by more than one group via variations of method, language/domains or data. Every group can determine their focus (within reason) themselves. When two groups use the same data, they can also work as if in a “competition”.

Topic Suggestions

1. Topic MarkertI: Semi-supervised learning for the automatic resolution of **metonymies**
2. Topic MarkertII: Improving **unsupervised sentence summarization and headline generation** with regards to fluency and fidelity
3. Topic MarkertIII: **Comparative anaphora resolution** as question answering (no slides, if interest will explain on blackboard)

Markert1: Semi-supervised Learning for the Resolution of Metonymies

> *"***Trope***: [...] jede Form der Rede, die das Gemeinte nicht direkt und sachlich durch das eigentl. Wort ausspricht, sondern [...] durch e. Anderes, Naheliegendes, e. "übertragenen" Ausdruck wiedergibt."*
>
> *Gero von Wilpert (1989): Sachwörterbuch der Literatur*

Frequent (every third sentence). Important for sentiment mining, text simplification, anaphora resolution, geographical IR ...

Examples

Metaphors

Use a similarity relationship between two domains (`ARGUMENT-IS-WAR`)

- He **attacked** my arguments.
- He **bashed** my arguments.

Metonymies

Use a contiguity relation between two domains (`PLACE-FOR-EVENT`)

- He was traumatized after **Vietnam**
- **Pearl Harbour** still has an effect on our foreign policy

Both types tend to be systematic and generalize over groups of words

Prior Work and Task

Most work focuses on metaphor resolution $\rightarrow$ this software project is metonymy recognition

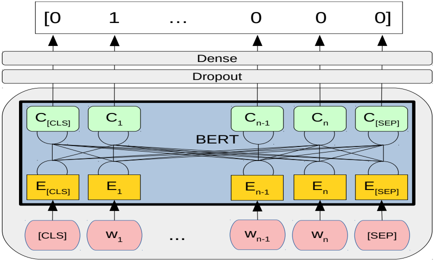
- He was traumatized after **Vietnam** $\rightarrow$ `PLACE-FOR-EVENT`
- **Brazil** lost the quarterfinal $\rightarrow$ `PLACE-FOR-TEAM`
- **Brazil** decided to stop deforestration $\rightarrow$ `PLACE-FOR-GOV`
- He lived in **Tokyo** $\rightarrow$ `LITERAL`
- **BMW** lost 3 points yesterday $\rightarrow$ `ORG-FOR-INDEX`
- He worked for **IBM** $\rightarrow$ `LITERAL`

Datasets

Dataset	Source	Type	Annot	literal	metos
Semeval-LOC[1]	BNC	Countries	Manual	1458	375
Semeval-ORG[2]	BNC	Companies	Manual	1211	721
ReLocar[3]	Wikipedia	Locations	Manual	995	1031
ConLL[4]	News	Locations	Manual noisy	4609	2448
WimCor[5]	Wikipedia	Locations	automatic	154322	51678

1, 2: Markert and Nissim, 2007
3, 4: Gritta et al., 2017
5: Mathews and Strube, 2020

State-of-the-Art: Li et al, 2020



Plus **masking of target word** in training and testing to avoid spurious information from rare target word occurrences:

He was traumatized by **Vietnam** $\rightarrow$ He was traumatised by **X**

Results Li et al (2020) (Accuracy)

Dataset	BL	BERT-BASE-MASK	BERT-LG-MASK
Semeval-LOC	80.1%	87.1%	88.2%
Semeval-ORG	62.7%	75.6%	77.2%
ReLocar	50.8%	93.9%	94.4%
ConLL	65.3%	93.7%	93.9%
WimCor	74.9%	95.4%	95.5%

This does not look too bad: what's the problem?

- Worst results on manually annotated datasets with diversity and **natural distribution**
- **Cross-domain accuracies** much lower: WimCor → Semeval 78.4% (worse than BL), WimCor → ReLocar 64.6%
- Ignores important target word information: **Vietnam** vs. **Solomon Islands** as `PLACE-FOR-EVENT`?

Greenland	4/100
Guyana	5/100
...	...
Japan	18/100
Hungary	21/100

But good target word info not easy to integrate with such small datasets

This does not look too bad: what's the problem?

- Worst results on manually annotated datasets with diversity and **natural distribution**
- **Cross-domain accuracies** much lower: WimCor $\rightarrow$ Semeval 78.4% (worse than BL), WimCor $\rightarrow$ ReLocar 64.6%
- Ignores important target word information: **Vietnam** vs. **Solomon Islands** as `PLACE-FOR-EVENT`?

Greenland	4/100
Guyana	5/100
...	...
Japan	18/100
Hungary	21/100

But good target word info not easy to integrate with such small datasets

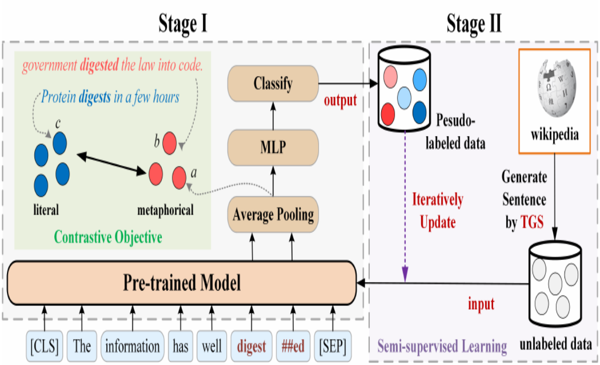
Semi-supervised learning for Figurative Language

Currently

Almost all work on metaphor or metonymy recognition is fully supervised. As especially the manually annotated metonymy datasets are small, this is a problem.

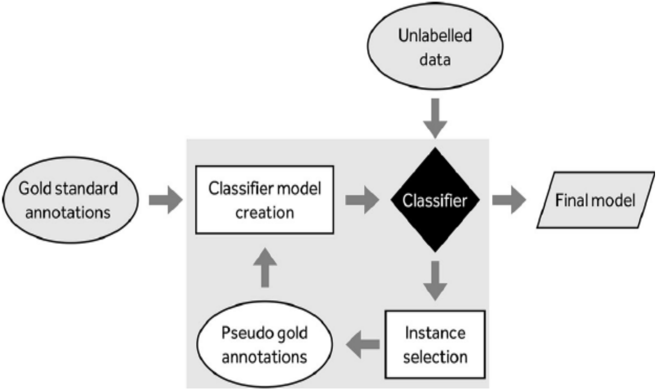
Recent exception for **metaphor**: CATE (Lin et al., EMNLP 2021): Use of self-training!

CATE's approach



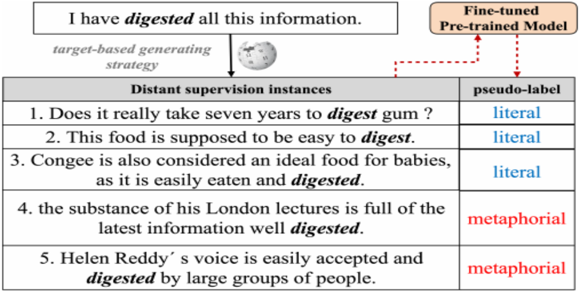
- Fine-tuning (and test) data: VUA metaphor corpus (BNC)
- Two contributions: contrastive objective (Stage I) plus self-training (Stage II)

Self-Training



Picture from Mihaila, C. and Ananniadou, S. (2014): *Semi-supervised learning of causal relations in biomedical scientific discourse.* In BioMedical Engineering Online.

Example for generated metaphor data in self-training



Distant supervision instances	**pseudo-label**
1. Does it really take seven years to ***digest*** gum ?	literal
2. This food is supposed to be easy to ***digest***.	literal
3. Congee is also considered an ideal food for babies, as it is easily eaten and ***digested***.	literal
4. the substance of his London lectures is full of the latest information well ***digested***.	metaphorial
5. Helen Reddy′ s voice is easily accepted and ***digested*** by large groups of people.	metaphorial

Picture from Lin et al (2021)

- Self-training has the problem of error propagation: CATE's solution is soft-labeling
- They show that self-training already helps for metaphor even without contrastive objective (simpler Stage 1)

Problems

- Only focuses on target word for dataset expansion, never the context
- Not used for metonymy
- No attempt to match labeled and unlabeled data domain (BNC $\neq$ Wikipedia)
- Only one semi-supervised paradigm

Markert I.1: Metonymy recognition with self-training

- Same self-training with soft labels on metonymies
- Expanded with domain matching (SemEval uses BNC unlabeled examples, Conll news etc).
- Include both context and target word in generating strategy:

  *He was traumatized by **Pearl Harbour***

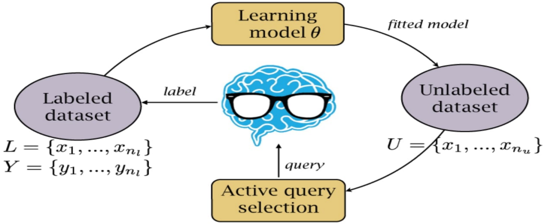
Target word-based

The attack on **Pearl Harbour**
The consequences of **Pearl Harbour**
. . .

Context-based

Americans had been traumatized by **Vietnam**
Traumatized by **Madrid**, Pochettino can't sleep anymore
. . .

Markert I.2: Metaphor/Metonymy Recognition with Active Learning



`deepai.org/machine-learning-glossary-and-terms/active-learning`

- Selection strategy crucial: often use examples where classifier is uncertain
- Advantage: added new data not noisy as human in the loop
- Can be simulated by holding out parts of training data as unlabeled data, if you don't want to annotate anything

Resources and Literature

- Lin et al (2021): *CATE: A constrastive Pre-Trained Model for Metaphor Detection with Semi-Supervised Learning*. In EMNLP 2021.
- Markert, K. and Nissim, M. (2007): *SemEval-2007 Task 08: Metonymy resolution at SemEval-2007.* In Semeval 2007.
- Markert, K. and Nissim, M. (2009): *Data and models for metonymy resolution.*. Language Resources and Evaluation, 43(2).
- Gritta et al. (2017): *Vancouver welcomes you! Minimalist location metonymy resollution.* ACL 2017.
- Mathews, K. and Strube, M. (2020): *A large harvested corpus of location metonymy*. In LREC 2020.
- Li et al (2020): *Target word masking for location metonymy resollution*. Coling 2020
- Ouali et al (2020): *An Overview of Deep Semi-supervised Learning.* `https://arxiv.org/pdf/2006.05278.pdf`

- All mentioned metonymy/metaphor data is publically available.

MarkertII: Sentence summarization/Headline Generation

The problem

Shorten a sentence or generate a headline from a news sentence, given a target length for the shortened sentence/headline

Example Pair

- ORIG: The word's biggest miner BHP Billiton announced Tuesday it was dropping its controversial hostile takeover bid for rival Rio Tinto due to the state of the global economy.
- HUMAN REFERENCE SUMMARY: BHP Billiton drops Rio Tinto takeover bid

Supervised vs. Unsupervised Methods

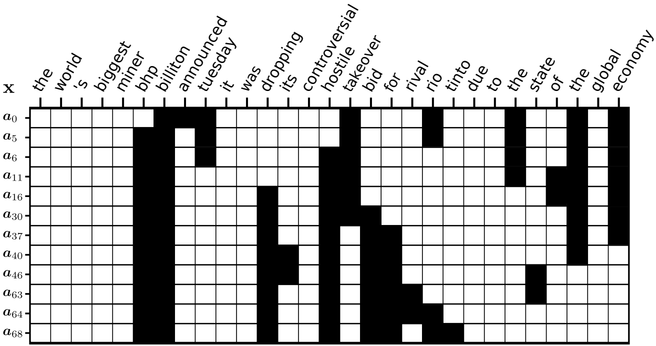
SUPERVISED

Many pairs given
Seq2Seq models

UNSUPERVISED

No pairs given
Source text maybe given
Target text maybe given

Schumann, Lou and Markert (ACL 2020): Unsupervised




```
summary:   bhp billiton dropping hostile bid for rio tinto
reference: bhp billiton drops rio tinto takeover bid
```

- Word-level Extraction
- Greedy Hill-climbing with restarts

Schumann et al: Objective Function

- Source Sentence $\mathrm{x} = (x_1, x_2, \ldots, x_n)$
- Output Sentence $\mathrm{y} = (y_1, y_2, \ldots, y_m)$
- $s < n$ summary upper bound
- Objective function f maximises for fluency and similarity

$$f(\mathrm{y}; \mathrm{x}, s) = f_{\overleftrightarrow{\mathrm{LM}}}(\mathrm{y}) \cdot f_{\mathrm{SIM}}(\mathrm{y}; \mathrm{x})^{\gamma} \cdot f_{\mathrm{LEN}}(\mathrm{y}; s), \tag{1}$$

- Fluency was measured via inverse perplexity of LSTMs trained on source or target sentences
- Similarity between y and x was measured by Sent2vec

Results

- State-of-the-art at the time for ROUGE score
- Human evaluation with 5 annotators via comparison to previous best models on 100 instances via fidelity and fluency

Models	Score (#wins/#ties/#loses)	
	Fidelity	**Fluency**
HC vs. WL	+0.18 (44/30/26)	+0.30 (45/40/15)
HC vs. ZR	+0.05 (35/35/30)	-0.03 (24/49/27)

Table: Human evaluation in a pairwise comparison setting on 100 headline generation instances.

Schumann et al: Example output

Good Example

- mubarak was ousted friday after being at the helm of his north african country for nearly 30 years .
- mubarak ousted after being at the helm of his country for years

Bad Example

- A third national security bill has been introduced to allow sharing of information between intelligence agencies and the Australian defence forces , allowing them to potentially target Australian terrorist fighters .
- bill introduced to allow sharing of information between intelligence agencies and terrorist

Schumann et al: Non-comparative performance analysis

- Preliminary annotation study of fluency and fidelity by Eric Kaiser (`http://misc.eric-kaiser.net/annotation`)
- 266 annotation
- Fidelity

Fidelity correct	39.9%
Fidelity incorrect	60.2%

- Fluency

1	18.8%
2	10.9%
3	9.8%
4	16.5%
5	44.0%

Markert II Project Ideas: Improve Fluency and/or Fidelity

- Better language models to improve fluency
- Use semantic graph matching methods (such as AMR scoring) as an objective function to improve fidelity

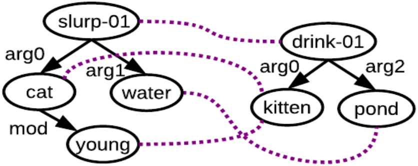


Figure 1: Similar AMRs, with sketched alignments.

Picture from Opitz et al. (2021)

Markert II Project Ideas: Improve Fidelity

The problem occurs in standard single-document summarization:

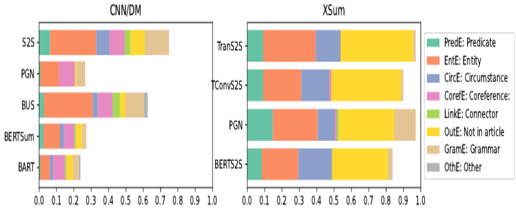


Figure from Pagnoni et al (2021) on 250 articles and their summaries

Idea

Adapt a suitable factual consistency evaluation metric from standard document summarization, such as FactCC (Kryscinscki et al (2019))

Ressources and Literature

- Raphael Schumann's code exists and runs
- Gigaword headline generation dataset: `https://github.com/harvardnlp/NAMAS`
- Schumann, Mou, Lu, Vechtomova and Markert (2020): *Discrete Optimization for Unsupervised Sentence Summarization with Word-level Extraction*. In ACL 2020.
- Kryscinski, McCann, Xiong and Socher (2020): *Evaluating the Factual Consistency of Abstractive Text Summarization*. In EMNLP 2020.
- Pagnoni, Balachandri and Tsvetkov (2021): *Understanding Factuality in Abstractive Summarization with FRANK: A benchmark for factuality metrics.* In NAACL 2021.
- Opitz, Daza and Frank (2021): *Weisfeiler-Leman in the Bamboo: Novel AMR Graph Metrics and a Benchmark for AMR Graph Similarity*. In TACL 2021.