Softwareproject Topics

Katja Markert: Topics WS22

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: Data augmentation for the automatic resolution of **metonymies**
2. Topic MarkertII: **Saints-Memory**: Matching saints in a German historical encylopedia to German Wikipedia

Markert1: Data Augmentation 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 → this software project is metonymy recognition

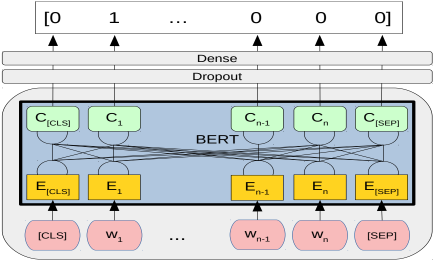
- He was traumatized after **Vietnam** → `PLACE-FOR-EVENT`
- **Brazil** lost the quarterfinal → `PLACE-FOR-TEAM`
- **Brazil** decided to stop deforestration → `PLACE-FOR-GOV`
- He lived in **Tokyo** → `LITERAL`
- **BMW** lost 3 points yesterday → `ORG-FOR-INDEX`
- He worked for **IBM** → `LITERAL`

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** → 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 $\rightarrow$ Semeval 78.4% (worse than BL), WimCor $\rightarrow$ ReLocar 64.6%
- **Overfitting** to Training Set
- Realistic datasets are **too small** and often **too unbalanced**

Current learning for Figurative Language

Currently

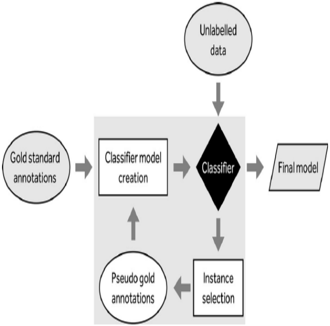
Almost all work on metaphor or metonymy recognition is fully supervised. As the manually annotated datasets are small, this is a problem.

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

Recent exception for **metonymy**: SWP Summer 2022: use of self-training with promising results

Semi-supervised learning vs. Data Augmentation

Self-Training



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

Data Augmentation: Add variations to input data (or feature-space) that are **label-preserving** or **predicably label-reversing**

By Backtranslation

IBM rose 4 points yesterday. → IBM stieg gestern um 4 Punkte. → IBM increased by 4 points yesterday.

Data Augmentation Possibilities

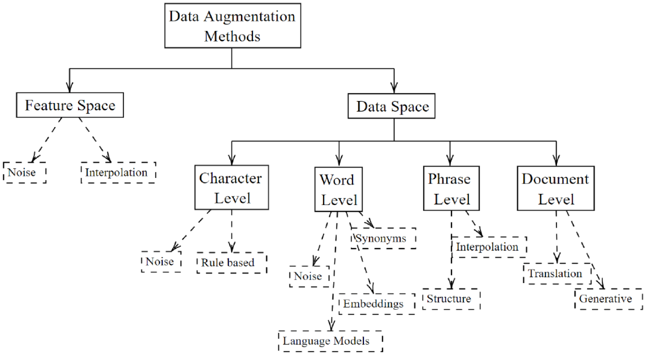


Figure 1: Taxonomy and grouping for different data augmentation methods.

Bild aus Bayer et al (2021): *A survey on data augmentation for text classification.* In: ACM Computing Surveys

Examples of data space data augmentation methods for metonymy recognition

Starting from *IBM rose 4 points yesterday.*

- By Backtranslation from/to German: *IBM increased by 4 points yesterday.*
- Grammar Variations: *IBM rises 4 points yesterday*
- Antonym or Synonym substitution: *IBM gained 4 points yesterday*
- Entity transformation: *IBM Corp. rose 4 points yesterday*
- Yoda Transformation: *Rose by 4 points yesterday, IBM did.*
- Noise transformations: *IBM rose 4 pints yesterday.*

Starting from *IBM shares rose 5 points yesterday*(literal) → *IBM rose 5 points yesterday* (metonymy)

Feature-space data augmentation

Interpolation with SMOTE or Mix-Up by interpolating, for example BERT layers:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

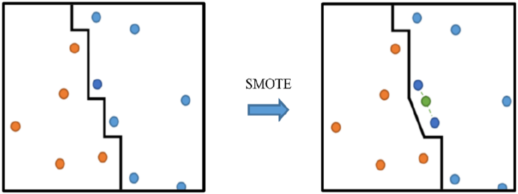


Figure 5: Illustration of the interpolation method SMOTE.

Bild aus Bayer et al (2021): *A survey on data augmentation for text classification.* In: ACM Computing Surveys

Advantages and challenges of Data Augmentation

- Not dependent on original classifier and its quality
- Many variations
- Tradeoff between meaning preservation and diversity, especially if your baseline model is a large language model

Challenges and Possibilities

- Identify label-preserving and predictably label-inverting transformations for metonymy
- Explore data augmentation strategies implemented in the NL-Augmenter (Dhole et al, 2021) to improve metonymy resolution
- Add own data augmentation strategies to the NL-Augmenter
- Enhance diversity of augmentation strategies
- Challenges:
  - Use knowledge gained in modules such as Statistical Methods, Syntax, Semantics to come up with sensible data augmentation strategies
  - Implement or re-implement such strategies
  - Integrate with a standard language model metonymy resolution baseline

Resources and Literature: data augmentation

- NL Augmenter: `https://arxiv.org/pdf/2112.02721.pdf` and the github page `https://github.com/GEM-benchmark/NL-Augmenter`
- Dhole et al (2021): *NL Augmenter: a framework for task-sensitive natural language augmentation.* `https://arxiv.org/pdf/2112.02721.pdf`
- Bayer et al (2021): *A survey on data augmentation for text classification.* In ACM Computing Surveys.

Resources and Literature: Metonymy

- 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 resolution*. In Coling 2020.

Saints-Memory

`https://encycnet.github.io/`: aims to create a new semantic resource for historical German in form of a knowledge graph.

- Currently 22 historic German encyclopedias
- Attempt to use DBSpotlight to automatically match entries to DBPedia (and GermaNet)

	Germanet	DBPedia
Brockhaus 1809	32.80	51.15
Eisler Philosophie 1904	46.06	25.40
Wander Sprichtwort 1867	43.75	16.06
Roell Eisenbahnen 1912	24.91	27.88
Heiligenlexikon 1858	2.34	0.35

Use cases:

- Finding gaps in Wikipedia
- Finding mismatched information

Heiligenlexikon

33,481 entries, 3m tokens

```
<b>S. Bilhildis</b>, (27. Nov.), Wittwe und Stifterin
des Klosters Altmünster (<i>Altum Monasterium B. V. M</i>.)
war die Tochter christlicher Eheleute von vornehmer Abkunft
Namens Iberius und Mechildis (Mechtildis, Mathildis) und
wurde zu Hochheim am Main um das Jahr 625 oder 626 geboren.
Was dieß für ein Hochheim am Main sei, ob der nicht
weit von Wirzburg gelegene Ort, gewöhnlich Veitshöchheim gen
...
Von ihrer Base zu Wirzburg in aller Gottseligkeit erzogen,
ward sie in jungen Jahren, etwa 16 oder 17 Jahre alt, an den
heidnischen Herzog Hettan (in Thüringen) vermählt,
...
Die Zeit, wann sie das Zeitliche
segnete, ist nicht zu ermitteln; Einige jedoch
setzen ihren Tod in das Jahr 630. (<i>El., Buc</i>.)
```


20 / 28

Bild

Bild von Joachim Schäfer - https://www.heiligenlexikon.de Ökumenisches Heiligenlexikon, Creative Commons CC BY-NC-SA 4.0

Wikipedia

Bilhildis von Altmünster, auch Bilihild, Bilehild oder Bilihilt und Bilhild (im 7. Jahrhundert in Veitshöchheim; gest. um 734 in Mainz) war eine fränkische Adelige, Klostergründerin und Äbtissin. Der Name Bilhildis ist althochdeutsch und bedeutet „die mit dem Beil Kämpfende". ...wurde sie gegen ihren Willen mit dem ungetauften, in Würzburg residierenden Herzog Heden (dux militum gentilis ... vocabulo Hetan) aus dem Geschlecht der Hedenen vermählt ...In Veitshöchheim findet jährlich an ihrem Gedenktag, dem 27. November, ein Gottesdienst statt.

That was the simple case...

- Unambiguous and successful name match in Wikipedia
- Matching of feast day ("Gedenktag")
- Even then we can see: name variations *Hettan - Heden*, different birth or death dates, uncertain information on very early saints

Normally...

- Obscure and unmatched saints:

```
<b>Boderius</b>, (22. Mai), wird in einigen
Orten als Martyrer verehrt
```

- Several names and several or ambiguous feast days
- Very ambiguous saints where name and day is not enough
  1. Bernardus (1): cistercian, abbot of Clairveaux, approx. 1100
  2. Bernardus (2): arch bishop of Vienne died 842, also named Barcar, or Barnar
  3. Bernardus (3): bishop of Carinola, died 1109
  4. ...
  5. Bernardus (64): simple monk of the Capucines, died 1540
- Normally no infoboxes in Wikipedia

Project Idea

- Define information extraction templates for saints

saint	relation	possible filler
saint	has-name	any string
saint	feast-day	date
saint	has-job	martyr, abbot, bishop, arch-bishop . . . ..
saint	is-born	date
saint	has-died	date
saint	located-in	location

- Algorithms for template filling inspired by IE work for Heiligenlexikon $\rightarrow$ Template 1
- Wiki API for approximate name match plus filtering
- Template filling from Wikipedia matches and Wikidata $\rightarrow$ Template(s) 2
- Template match
- Evaluation

Template filling

- There is no gold data, so unsupervised and probably no or little standard machine learning
- Programming from scratch
- Possibility:
  1. Preprocessing with Heideltime and German Stanford Core NLP
  2. Template 1: Some relations can be filled by (approximate) regular expression match and tag restrictions (names, job titles, feast days)
  3. Template 2: REs or Wikidata `https://www.wikidata.org/wiki/Q477895`
  4. Perform simple, unambiguous saint matches
  5. Bootstrapping for natural language patterns from these seeds: iterative algorithm
  6. Enhanced with semantic similarity of texts

Extensions

- Some relations are shared with standard IE on news. Use existing algorithms trained on news for domain transfer.
- Use additional resources such as Wikipedias in different languages or `https://www.heiligenlexikon.de/`

Resources and Literature

- Encyc-Net for the Heiligenlexikon: `https://encycnet.github.io/`
- Hagen et al (2020): Twenty-two historical encyclopedias encoded in TEI: A new resource for the digital humanities. In: LaTeCH-CLfL 2020: 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.
- Jurafsky and Martin (2022): Speech and Language Processing, 3rd edition. Chapter 17.
- For the TACRED news IE relation dataset: Zhang et al (2017). Position-aware attention and supervised data improve slot filling. In *EMNLP 2017*.
- Stoica et al (2021). Re-tacred: Addressing shortcomings of the tacred dataset. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 15. 2021.
- Han et al (2021): PTR: Prompt Tuning with rules for text classification. `https://arxiv.org/pdf/2105.11259.pdf`
- Peters et al (2019). Knowledge Enhanced Contextual Word Representations. In *EMNLP 2019*