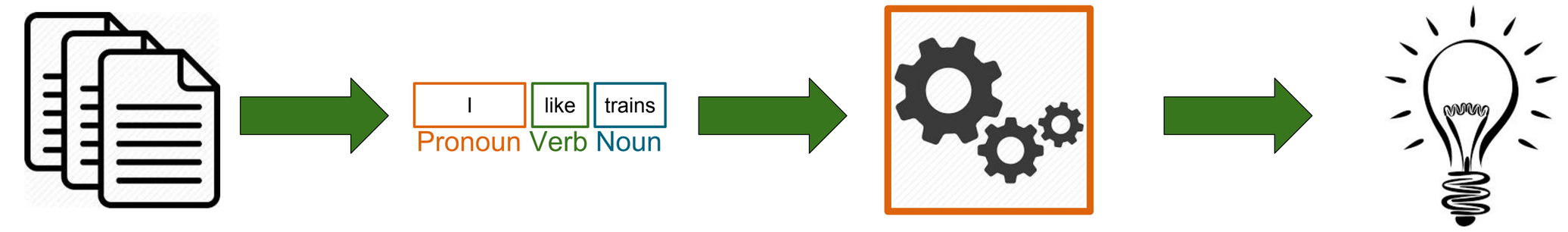
Word Sense Disambiguation

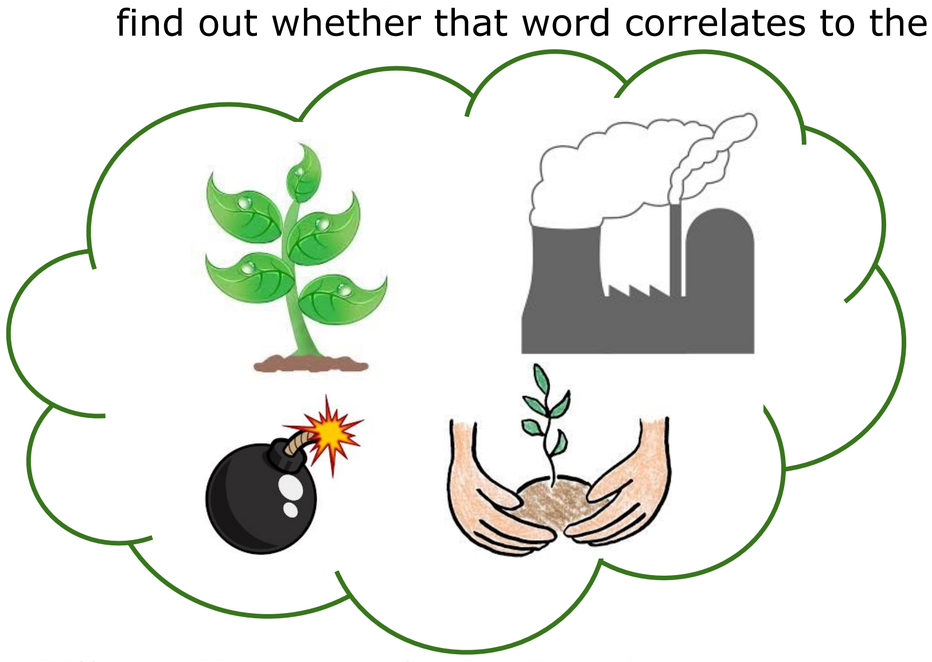
Find correct word senses in text

Studies[1] have shown that 11.5% of the overall vocabulary in a text is ambiguous, even 40% in English prose. In practical text processing applications that may cause false assumptions about the data and creating worse results than need to be. Word Sense Disambiguation is a research field in Natural Language Processing that concerns the determination of the correct sense of a word in a given context.



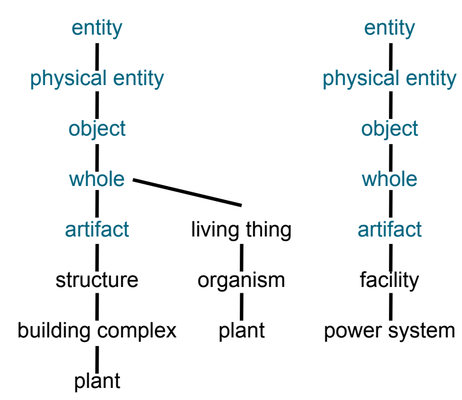
Text Documents | Part-Of-Speech Tagging | Word Sense Disambiguation | Improved semantic knowledge

Example: Imagine the SAP Service Ticket Intelligence wants to evaluate a ticket with the word “plant” in it and wants to find out whether that word correlates to the biological or industrial sense to create a reply according to the correct sense.



Selectional Association Algorithm

Resnik’s algorithm[2] uses a similarity measure between the ambiguous word and the word most associated to the current context given as input.
At first the algorithm checks the WordNet entries of all possible senses of both words to get their taxonomy trees. The more they overlap the higher the similarity measure is.
Unfortunately this algorithm relies on to many assumptions concerning the correctness of the association finding, Wordnet dataset and the dataset.



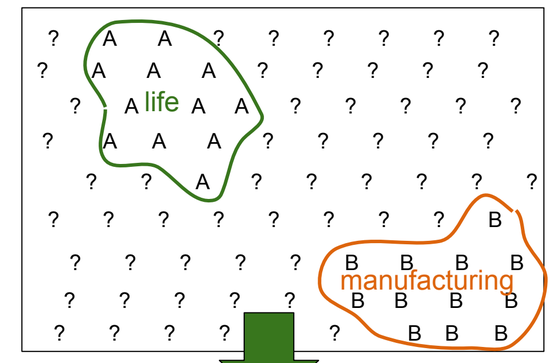
Comparison of the association ‘power system’ with the ambiguous word plant

Different Meanings of “plant” in WordNet:

- <noun.artifact>: works, industrial plant (buildings for carrying on industrial labor)
- <noun.Tops>: flora, plant life ((botany) a living organism lacking the power of locomotion)
- <noun.person>: an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience
- <noun.cognition>: something planted secretly for discovery by another
- <verb.contact>: set (put or set (seeds, seedlings, or plants) into the ground)
- <verb.contact>: implant, engraft, embed, imbed, fix or set securely or deeply
- <verb.creation>: establish, found, constitute, institute (set up or lay the groundwork for)
- <verb.possession>: place into a river
- <verb.contact>: place something or someone in a certain position in order to secretly observe or deceive
- <verb.cognition>: implant (put firmly in the mind)

Semi-Supervised Machine Learning

To make less assumptions about the input this task was introduced to the Machine Learning field. The main problem for supervised learning of this task is that training data is scarce. That’s why Yarowsky’s algorithm[3][4] proposes a bootstrapping semi-supervised learning. During the training the most confident predicted words for a sense are added to the training set for the next iteration.



Collocational Features

Original POS Tagging	The	scientist	examined	the	**plant**	with	microscopic	care.
	DT	NN	VBD	DT	NN	IN	NNS	NN

3 word window

Co-occurrence Features

stemmed filtered words

Sense A: {scientist: 0, examine: 0, microscopic: 1, care: 0}
Sense B: {scientist: 0, examine: 0, microscopic: 0, care: 0}
frequency of appearing in example dataset for this sense

Use Cases

XING:
identify business domain of description text
SAP:
Service Ticket Intelligence
intelligent Chat Bots
Amazon:
identify product class of product description
Other:
Machine Translation

State of Research

The problem of finding the correct sense of a word remains AI-complete after all, that means computers still aren’t as good as the human brain in this domain. But most of the presented and researched algorithms look promising as they are at least better than the baseline of just choosing the most frequent sense for a word. There are more ideas in solving this problem like giving partial credit to the machine learning at least the right domain in a field or considering the occurences of the same word sense in a bigger context (“one sense per discourse”).
Multiple algorithms are available in Frameworks like NLTK (Python)[5] that hopefully improve many Natural Language Processing applications in the future.

[1] DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. Computational Linguistics,14, 31–39
[2] Resnik, P. (1998). Wordnet and class-based probabilities. In Fellbaum, C. (Ed.), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA
[3] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In ACL95, Cambridge, MA, pp. 189–196. ACL
[4] Hearst, M. A. (1991). Noun homograph disambiguation. In Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research, Oxford.
[5] http://www.nltk.org/howto/wsd.html
Overall:
Jurafsky, D. and J. H. Martin (2000) *Speech and Language Processing 1st Edition*
Jurafsky, D. and J. H. Martin (2017) *Speech and Language Processing 3rd Edition draft*

Maxi Fischer

Master student
Hasso Plattner Institute, Potsdam, Germany

E-Mail: maxi.fischer@student.hpi.de