Improving SMT-based Synonym Extraction across Word Classes by Distributional Reranking of Synonyms and Hypernyms

Universität Stuttgart

Maximilian Bräuninger **Marion Weller-Di Marco** **Sabine Schulte im Walde**

Goals

- **extract synonym candidates** using parallel corpora and word alignment

- **rank synonym candidates** according to translation probabilities (see example)

- **improve synonym probability rating** using reranking based on **hypernym detection** and **semantic similarity**

Gold Standard

- synonym candidates extracted from the online **German Dictionary *Duden***

- **broad amount of synonyms** compared to other sources

- **manual evaluation** indicates gold standard might have **problems with ambiguous words**

Target Sets

- **3 different target sets** for **nouns** (NN), **verbs** (VV) and **adjectives** (ADJA)

- **300 words per set** split into high-, medium- and low-frequency words evenly

- **words with high and medium frequency** achieve **better results** than words with low frequency

Results

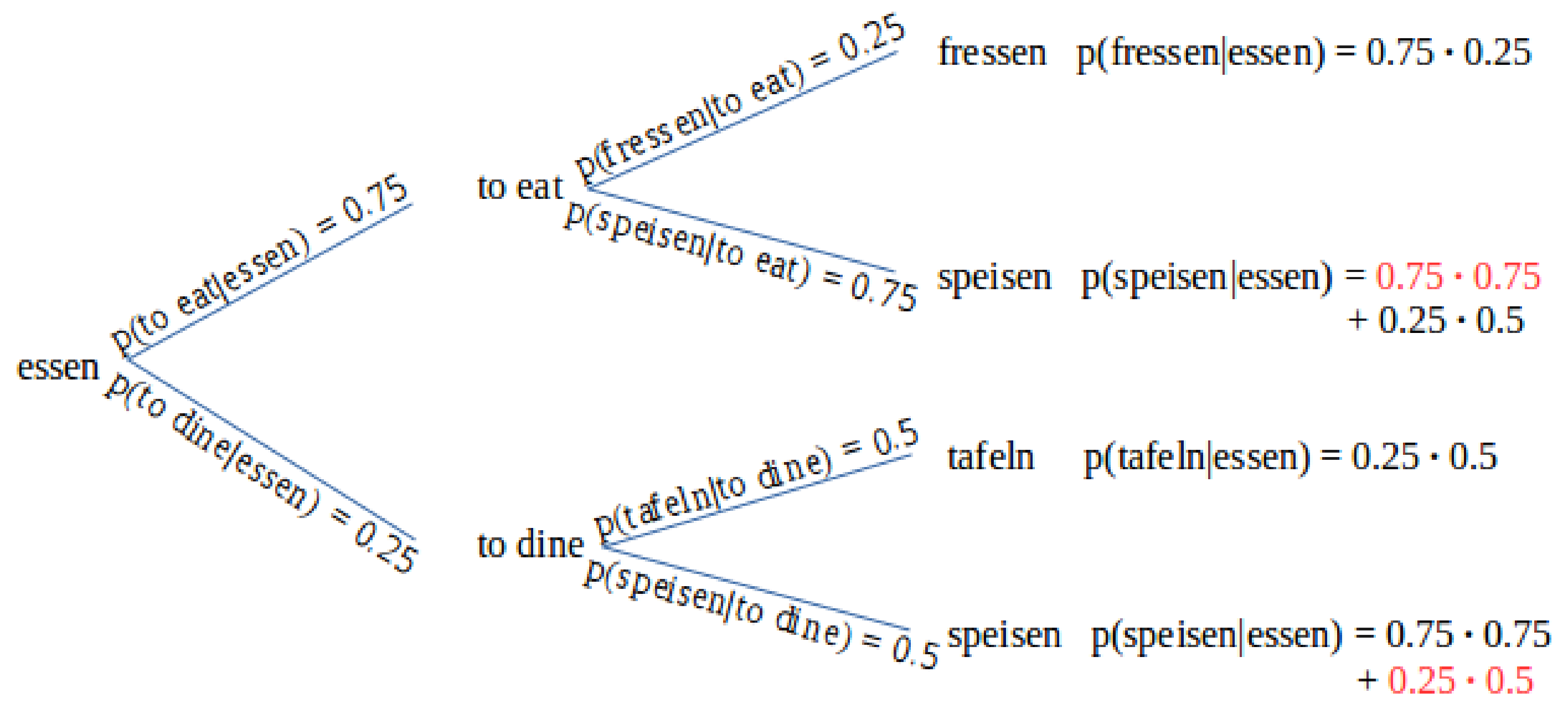
- **adjective synonyms** have **highest precision**

- **precision value decreases at 5 and at 10**

- **reranking techniques** show **no improvement**

		Precision at 1	Precision at 5	Precision at 10
adjectives	unranked	**62%**	**42%**	**33%**
	cosine similarity	**62%**	**42%**	**33%**
	weeds precision	46%	37%	32%
nouns	unranked	48%	**33%**	**25%**
	cosine similarity	**49%**	32%	**25%**
	weeds precision	30%	24%	21%
verbs	unranked	**44%**	**32%**	**25%**
	cosine similarity	41%	30%	24%
	weeds precision	32%	26%	23%

Example


essen
p(to eat|essen) = 0.75
p(to dine|essen) = 0.25
to eat
p(fressen|to eat) = 0.25
p(speisen|to eat) = 0.75
to dine
p(tafeln|to dine) = 0.5
p(speisen|to dine) = 0.5
fressen p(fressen|essen) = 0.75 · 0.25
speisen p(speisen|essen) = 0.75 · 0.75 + 0.25 · 0.5
tafeln p(tafeln|essen) = 0.25 · 0.5
speisen p(speisen|essen) = 0.75 · 0.75 + 0.25 · 0.5

Reranking Math

- **2 different** approaches using a vector space model

- **weeds precision** value detecting **hypernyms**

$$\mathrm{weedsPrecision}(u,v) = \frac{\sum_{f\in(F_u\cap F_v)} w_u(f)}{\sum_{f\in F_u} w_u(f)}$$

- **cosine similarity** ranking **semantic similarity** of two words

$$\mathrm{cosineSimilarity}(u,v) = \frac{\sum_{f\in(F_u\cap F_v)} w_u(f)\cdot w_v(f)}{\sqrt{\sum_{f\in F_u} w_u(f)^2}\cdot\sqrt{\sum_{f\in F_v} w_v(f)^2}}$$