Cross-lingual bootstrapping of Semantic Lexicons: The case of FrameNet

Sebastian Padó
Comput. Linguistics
Saarland University
Germany

Mirella Lapata
School of Informatics
University of Edinburgh
UK

Motivation

- Long-term goal: Predicate-argument structure for free text (helpful for IE, IR, QA)
- Prerequisite: **Semantic lexicon (this study)**
  - Generalise over individual predicate
  - Semantic roles may be sense-specific
- Lexicons available almost exclusively for English
  - Manual construction very costly
- Key idea: **Bootstrap** semantic lexicon for new language by exploiting existing English lexicon
  - Project English entries using **parallel** corpus

Outline

- Semantic Lexicons
  - FrameNet
- Cross-lingual bootstrapping
  - Method
  - Filtering
- Experiments
- Conclusions

Semantic Lexicons

FrameNet

Semantic Lexicon for English, based on "Frame Semantics" (Fillmore 1984)

Frame: COMMITMENT		
Frame Elements	SPEAKER	**Kim** promised to be on time.
	ADDRESSEE	Kim promised **Pat** to be on time.
	MESSAGE	Kim promised Pat **to be on time.**
	TOPIC	The government broke its promise **about taxes.**
	MEDIUM	Kim promised **in writing** to sell Pat the house.
FEEs	consent.v, covenant.n, covenant.v, oath.n, vow.n, pledge.v, promise.n, promise.n, swear.v, threat.n, threaten.v, undertake.v, undertaking.n, volunteer.v	

FrameNet

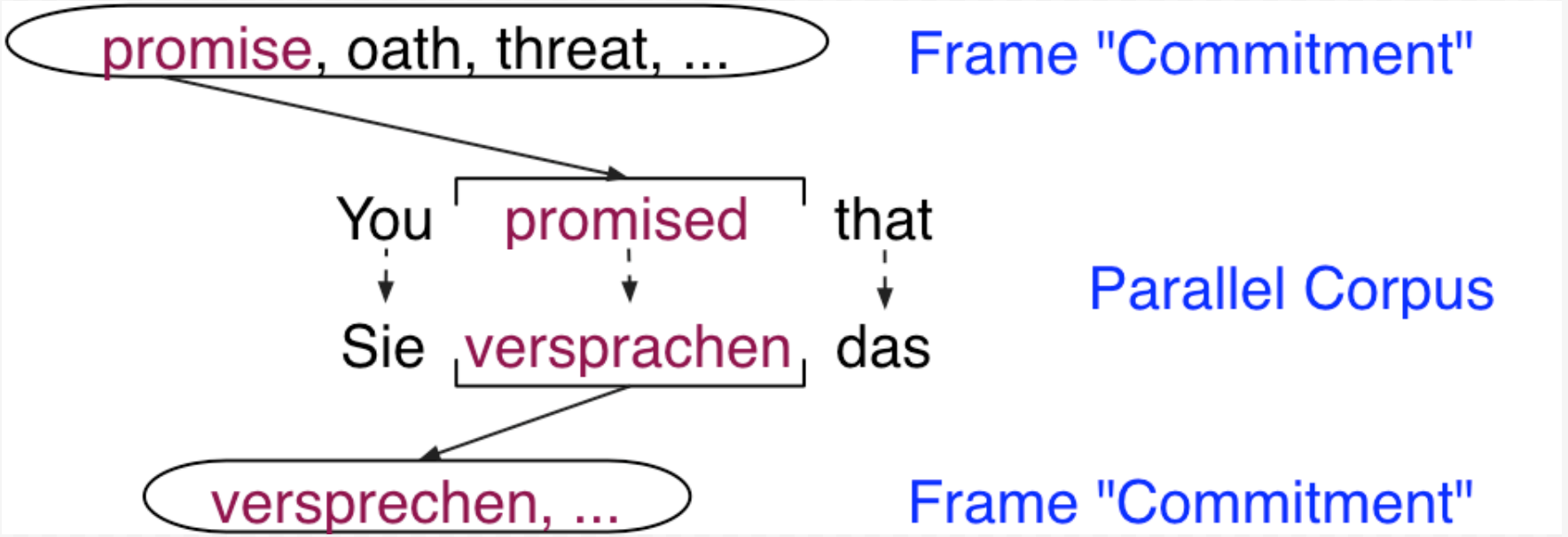
- Ongoing development at ICSI (UC Berkeley)
- Size: > 500 Frames, > 7.000 Lemmas

- FrameNet lexicon structure is (largely) **language-independent**
  - Reason: Reference to conceptual situations
  - Consequence: **Frames** should generalise across languages (German, Spanish, Japanese, …)
    - Only need to identify new **lemmas**

Cross-lingual Bootstrapping

Projection in parallel corpora

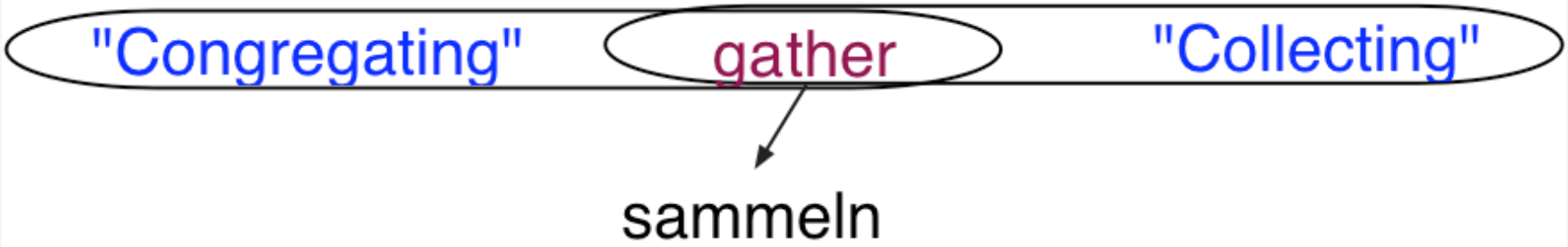
Naïve Method: Use word alignments in a large parallel corpus; assume that pair members belong to same fame



Projection Errors

- **Problem 1:**
  Word alignment

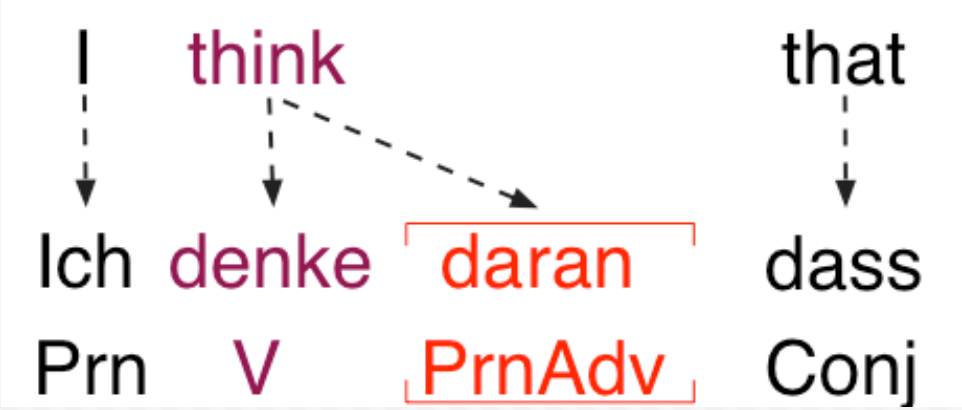
- **Problem 2**: Frame polysemy



- **Solution: Generate and prune**
  - **Generate** all possible translation pairs
  - **Prune** translation pairs by applying **shallow** filters

Token-level Filters

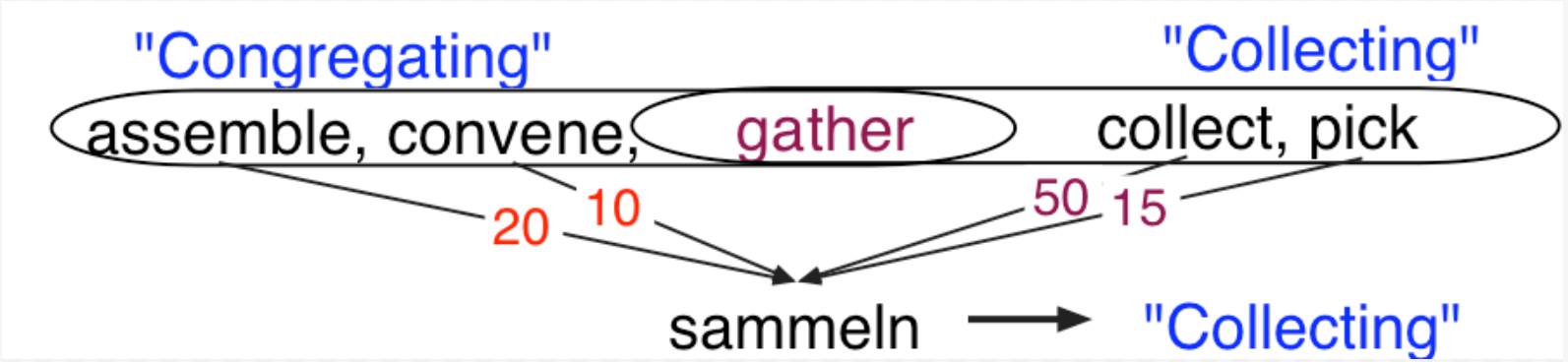
- Discard unreliable alignment **tokens**
- Filter 1:
  Content words only



- Filter 2: Alignment Intersection
  (Koehn, Och, and Marcu 2003)

Type-level filters

- Discard unreliable **translation pairs**
- Filter 1: “Most frequent frame” heuristic



- Filter 2: “Low translation entropy” heuristic (Modification of Melamed 1997)

daran: believe (5), remind (4), think (2), …	unreliable
erinnern: remind (25)	reliable

Experiments

Experimental Setup

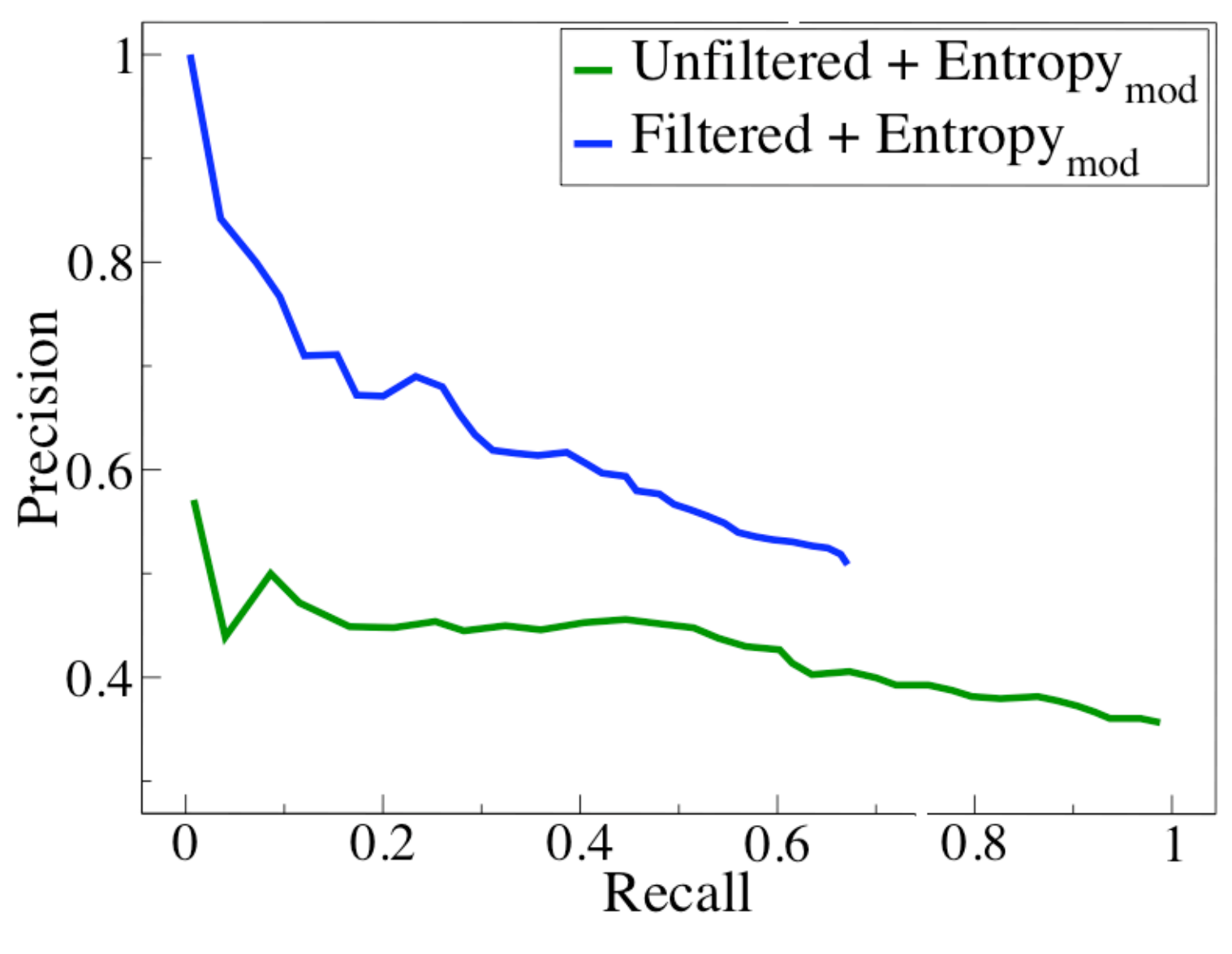
- EUROPARL corpus (Koehn 2002)
  - English ⇒ German, English ⇒ French
  - Automatic word alignment with GIZA++ (Och and Ney 2003)
- Apply method to complete FrameNet lexicon
- Manual evaluation for 15 frames
  - Random sample from three frequency bands
  - Double annotation, Kappa 0.78/0.79

German: Token level filters

Model	Rec.	Prec.	F
Unfiltered	1.00	0.35	0.52
Bidir. Alignment	0.70	0.47	0.56
Content words	0.98	0.40	0.57
Both	**0.68**	**0.50**	**0.58**

Simple token level filters improve precision and F-score

German: Type level filters



Type-level filters: still higher precision (**seed lexicon**)

Conclusion

- Knowledge-lean method to acquire FrameNet lexicons for new languages
  - Exploit word alignments in parallel corpus
  - Generate-and-prune strategy
  - Result: high-precision lexicon
- Current research: Automatic projection of semantic roles