POS Induction with Distributional and Morphological Information Using a Distance-Dependent Chinese Restaurant Process

Kairit Sirts[1], Jacob Eisenstein[2], Micha Elsner[3], Sharon Goldwater[4]

TALLINN UNIVERSITY OF TECHNOLOGY 1
INSTITUTE OF CYBERNETICS
Georgia Institute of Technology 2
4 THE UNIVERSITY of EDINBURGH informatics
ilcc Institute for Language, Cognition and Computation
3 THE OHIO STATE UNIVERSITY

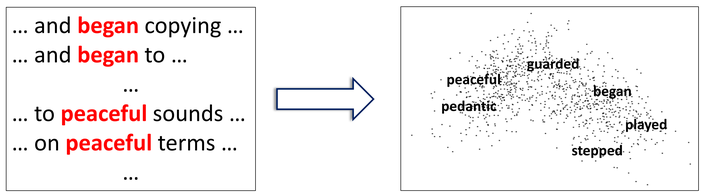
Clustering for POS Induction

We use a distance-dependent Chinese restaurant process [1] (dd-CRP) to incorporate **distributional** and **morphological** information for POS induction.

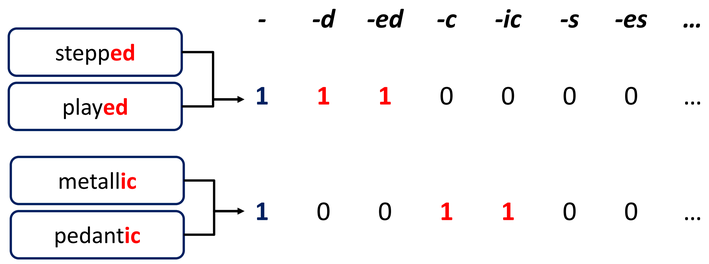
- Distributional info modelled as Gaussian likelihood over **word embeddings**.
- Morphological info modelled in prior using feature-based suffix similarity function.
- Learn the number of fine-grained morphosyntactic clusters.
- Learn the suffix similarity function for the dd-CRP prior.
- Tested on English, better performance than K-means and infinite Gaussian mixture model (IGMM); state-of-the-art for fine-grained POS induction.

Distributional and Morphological Data

Word embeddings represent the contextual information with continuous vectors trained using neural networks. We use the Polyglot vectors (pretrained on Wikipedia) [2].



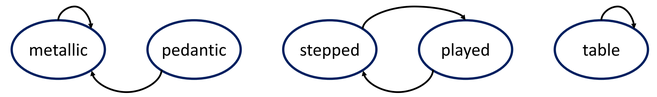
Morphological information is represented using suffix-based feature vectors.



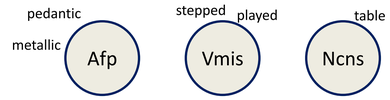
IGMM with dd-CRP Prior

Distance-dependent Chinese restaurant process:

dd-CRP is a variant of a regular CRP, where each customer **follows** another customer with probability proportional to the **distances** between the customers.

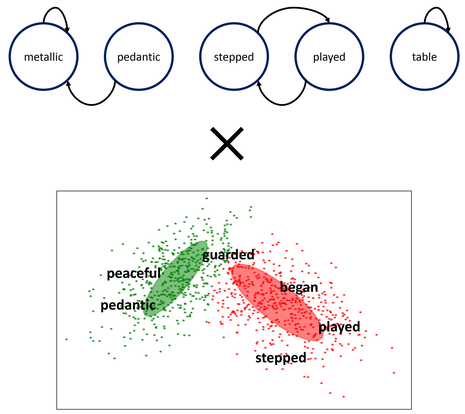


$$P(\text{stepped} \rightarrow \text{played}) \propto e^{w^T f(\text{stepped, played})}$$
$$P(\text{table} \rightarrow \text{table}) \propto \alpha$$



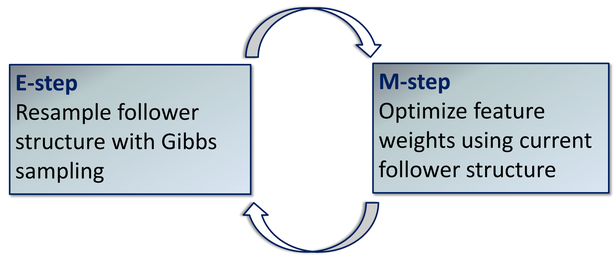
Model:

We combine the ddCRP prior with Gaussian likelihood and model distances using the suffix feature function $f(\text{word1}, \text{word2})$.



Inference:

We use a variant of Monte Carlo Expectation-Maximization:



Experiments on Multext-East English

Tagset:

Fine-grained tags (104) are sequences of morphological features:

Afp – **A**djective **q**ualificative **p**ositive (*metallic, pedantic*)

Vmis – **V**erb **m**ain **i**ndicative pas**t** (*stepped, played*)

Ncns – **N**oun **c**ommon **n**euter **s**ingular (*table*)

Evaluation:

Evaluation is **type-based**; this takes low-frequency words more into account.

Baselines:

- **K-means** – uses only word embeddings with K = gold number of clusters
- **IGMM** – uses word embeddings and learns the number of clusters

Three models using dd-CRP:

- **Uniform** distance function – almost identical to IGMM
- **Learned** distance function – log-linear feature model
- **Exponentiated** – learned prior is emphasized by an exponent parameter a = 5

				K-means	
Model	**K**	**1-1**	**V-m**	**1-1**	**V-m**
K-means	104	16.1	47.3	-	-
IGMM	55.6	41.0	45.9	23.1	49.5
dd-CRP uniform	80.4	50.5	52.9	18.6	48.2
dd-CRP learned	89.6	50.1	55.1	17.6	40.0
dd-CRP exp	47.2	**64.0**	**60.3**	25.0	50.3

References

[1] D. M. Blei and P. I. Frazier. 2011. Distance dependent Chinese restaurant process. *Journal of Machine Learning Research*

[2] R. Al-Rfou et al. Distributed word representations for multilingual nlp. In *CoNLL'13*

Acknowledgements

Supported by the Tiger University Program of the Estonian Information Technology Foundation for Education for the first author.

tiigriülikool