Application of Word Embeddings in Cross-Lingual Word Sense Disambiguation

Mitchell Stern
Advised by Dr. Lyle Ungar, Department of Computer and Information Science

Problem Description

Given a polysemous word in context…

- select a sense label from a manually collated sense inventory. (Monolingual)
- produce a suitable translation into various target languages. (Cross-Lingual)

Example: The nuclear power plant produced energy for the entire country.

Monolingual – A factory or workshop for the manufacture of a particular product.
Cross-Lingual – Spanish: planta, Dutch: kerninstallatie, German: atomkraftwerk, Italian: impianto, French: centrale

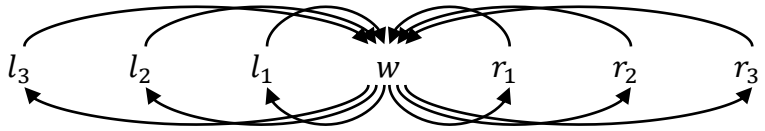
System Design

Lexical features

- Words, lemmas, parts of speech, combinations thereof
- Generated for a context of 3 words on either side

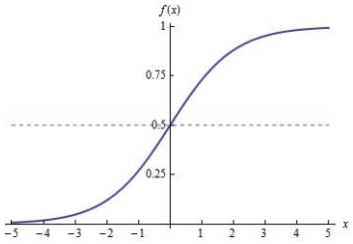
Word embeddings

- Dense, low-dimensional, real-valued vectors
- Capture syntactic and semantic information
- Induced using neural networks or by finding spectral decompositions of word co-occurrence matrices



Maximum entropy classifier

- Uses multinomial logistic regression
- Predicts a probability for each translation



- Functional form: $f(x) = 1/(1 + e^{-x})$
- Trained using limited-memory BFGS

Results

Our system was trained and tested on the same data used in the SemEval 2013 cross-lingual WSD competition, achieving highly competitive results for the BEST metric and state-of-the-art results for the OUT-OF-FIVE metric. The scores below indicate translation accuracy on a scale of 0 to 100.

BEST	Spanish	Dutch	German	Italian	French
Baseline	23.23	20.66	17.43	20.21	25.74
Preliminary	28.67	21.37	20.64	23.34	27.75
Spectral HMM	27.56	23.87	21.30	22.50	28.84
C&W	29.76	**25.17**	22.47	23.59	30.20
HLBL	28.34	24.60	22.35	23.13	29.54
Word2Vec	29.59	25.07	22.74	23.64	30.23
LR-MVL	30.72	25.11	22.97	24.85	30.73
One-Step CCA	30.76	24.63	**23.17**	24.71	30.69
Context-Specific	30.76	24.68	23.02	24.64	30.65
Two-Step CCA	30.76	24.99	23.08	24.86	30.77
OSCCA (Large)	31.15	24.82	23.03	25.12	**30.92**
SemEval Best	**32.16**	23.61	20.82	**25.66**	30.11

OUT-OF-FIVE	Spanish	Dutch	German	Italian	French
Baseline	53.07	43.59	38.86	42.63	51.36
Preliminary	60.93	46.12	43.40	51.89	57.91
Spectral HMM	61.00	47.86	44.15	53.88	60.35
C&W	62.07	48.81	45.06	55.42	63.21
HLBL	61.11	47.25	44.51	55.16	61.19
Word2Vec	62.32	48.74	45.51	56.04	62.64
LR-MVL	63.11	49.63	47.03	56.40	**63.51**
One-Step CCA	63.08	49.85	46.60	56.36	62.98
Context-Specific	**63.67**	48.86	**47.05**	56.18	62.85
Two-Step CCA	63.04	**49.98**	46.75	55.91	63.37
OSCCA (Large)	62.95	49.91	46.39	**56.65**	63.35
SemEval Best	61.69	46.55	43.66	53.57	59.80