WORD SENSE INDUCTION USING DOMAIN-SPECIFIC CORPORA

James A. Boyden, James R. Curran
ə-lab, School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

When we talk about a **mouse**, do we mean a rodent or a computer device?
When we talk about a **star**, do we mean an actor or a point of light in the sky?

Words can have multiple *senses*; without knowing which sense of a word was intended, we can't interpret the overall **meaning**.
Word sense induction is the task of discovering the senses of a word that occur in a given corpus of text.

NEED FOR WORD SENSE INDUCTION

- Manually-constructed word sense inventories exist, but they have problems:
  1. missing senses
  2. out-of-date senses
  3. senses that are too fine-grained for practical usage
     (see Figure 1 for an example)

1.	*a celestial body of hot gases that radiates energy derived from thermonuclear reactions . . .*
2.	*someone who is dazzlingly skilled in any field*
3.	*any celestial body visible (as a point of light) from the Earth at night*
4.	*a plane figure with 5 or more points; often used as an emblem*
5.	*an actor who plays a principal role*
6.	*a performer who receives prominent billing*
7.	*a star-shaped character * used in printing*
8.	*the topology of a network whose components are connected to a hub*

Figure 1: *The 8 senses of* `star` *given in WordNet 2.0*

DISTRIBUTIONAL SIMILARITY

- Distributional similarity can calculate the similarity of the **meanings** of words, based upon the **contexts** in which the words appear [Lin, 1998]:

I fed the { ⋮ *dog* ***mouse*** *pigeon* *snake* ⋮ } .
object of the verb "feed"

The { ⋮ *dog* ***mouse*** *child* *thief* ⋮ } ran away.
subject of the verb "run"

- Each context extracted from the corpus becomes a dimension in a high-dimensional **vector space**.
- Each distinct word in the corpus becomes a vector.
- These vectors can be compared (*cosine*, Jaccard, *etc.*) to calculate the similarity of word meanings.

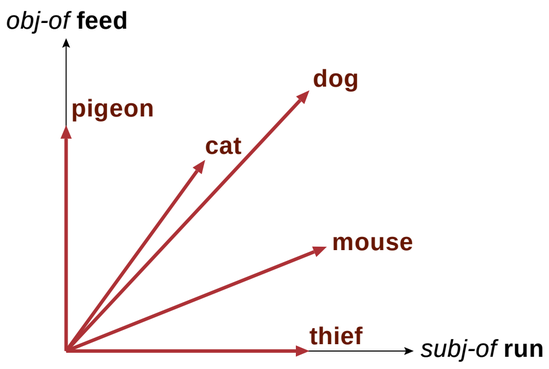

Figure 2: *A hypothetical 2-D distributional similarity example*

- A corpus might yield **tens of millions** of contexts.

Word Sense Confusion

- However, distributional similarity does not distinguish between different senses of a word.
- The contexts of different senses will be combined:

{ feed, catch, trap } the ***mouse*** { click, drag, move } the ***mouse***
Combined contexts of the different senses of `mouse`

USING DOMAIN INFORMATION

- Particular word senses often predominate in specific topic domains [Gale et al., 1992, Magnini et al., 2002, Koeling et al., 2005] — *e.g.* `mouse` usually means:
  – a rodent when in the BIOLOGY domain
  – a computer device when in the COMPUTING domain

OUR RESEARCH QUESTION

If we extract contexts for a word from domain-specific corpora, and calculate distributional similarity for that word within each domain, will we be able to distinguish **domain-specific senses**?

DOMAIN-SPECIFIC CORPORA

- We devised **49 high-level topic domains** (*e.g.* HISTORY, PHYSICS, MUSIC, SPORT, LAW, TRANSPORT, FOOD+COOKING, FURNITURE, RELIGION, CLOTHING+FASHION, BANKING+FINANCE, SEXUALITY)
- For each domain, we created a large domain-specific text corpus by downloading topic-specific pages from the Dmoz web directory.
- We extracted contexts for each corpus.

	Words	Contexts
max	1,669 M	224 M
mean	301 M	63 M
min	47 M	12 M

Table 1: *Per-corpus word-count and context-count statistics*

DOMAIN-SPECIFIC THESAURI

- Given a set of contexts extracted from a corpus, the `thesaurus` program [Curran, 2004] uses distributional similarity to calculate the **200 closest synonyms** to a given query word.
- For each of our **500 evaluation words**, we calculated the synonyms for that word in each domain, obtaining **49 domain-specific thesauri**.

Astronomy		Acting Performance	
stars	0.1124	stars	0.07593
object	0.04925	starring	0.05141
galaxy	0.04866	actor	0.03776
planet	0.04783	starred	0.02877
sun	0.03819	played	0.02756
galaxies	0.03658	opposite	0.02756
sky	0.0353	actress	0.02752
cluster	0.03418	plays	0.02654
moon	0.03186	like	0.02429
objects	0.03091	director	0.02381
Music		**Ball/Racquet Sports**	
stars	0.06246	stars	0.05229
singer	0.03685	rookie	0.03379
artist	0.0309	veteran	0.03362
legend	0.02796	player	0.0304
superstar	0.02678	champion	0.03015
songwriter	0.02467	superstar	0.02923
featuring	0.02236	senior	0.02887
musician	0.02219	junior	0.02774
hits	0.0215	midfielder	0.02686
friend	0.02137	legend	0.02653
Visual Art		**Craftsmanship**	
stars	0.04145	stars	0.05747
crescent	0.01957	heart	0.02433
hiking	0.01897	flower	0.02139
flower	0.01722	needles	0.02056
heart	0.01628	square	0.01992
oval	0.0159	pieced	0.01958
anchor	0.0145	diamond	0.01947
sun	0.01448	leaf	0.01851
cross	0.014	pinwheel	0.01726
square	0.01327	block	0.01688

Figure 3: *Some domain-specific synonyms for* `star`*. The numbers are in-corpus vector-similarity values.*

COMPARING SYNONYM LISTS

- We also combined all domain-specific corpora into a single **All-Domains** corpus, which gives an approximation for "general" (non domain-specific) text.
- For each word, we compared each domain-specific synonym list with the All-Domains synonym list to discover when a word has **domain-specific senses**.
- We calculated the similarity of pairs of synonym lists using **Rank Biased Overlap** [Webber et al., 2010].

0.54	music
0.38	acting-performance
0.37	ball+racquet-sports
	⋮
0.30	astronomy
	⋮
0.01	engineering
0.00	hunting-sports
0.00	chemistry

Figure 4: *Some synonym-list similarity scores for* `star`

- Figure 4 displays some RBO ($p = 0.92$) similarity scores, comparing each domain-specific synonym list to the All-Domains synonym list for `star`.
  – Scores lie in the range $[0, 1]$, where 1 indicates identical lists.
  – The relatively low "top score" of 0.54 indicates that `star` has no single predominant sense; rather, general text contains a mixture of several senses.

EXPERIMENTS & RESULTS

- Given a query word, we want our system to retrieve **all domains with domain-specific senses**.
- We evaluate system results against a **gold standard** (the WordNet Domains [Bentivogli et al., 2004] of the word).
- We calculate **Precision** & **Recall** of the results.
- We vary the system parameters (such as RBO p-value, synonym-list similarity thresholds and clustering algorithm) to obtain the **best average F_1-measure** over all words.

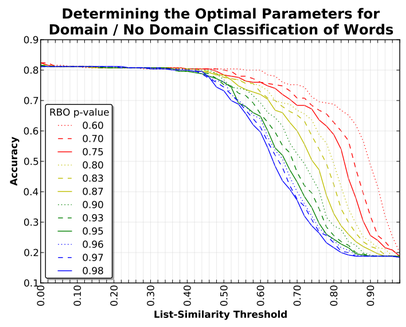

Figure 5: *Results of an experiment to partition words into Has Domains vs. Has No Domains classes*

CONCLUSION

- Initial results are encouraging, but missing domains in the gold standard are problematic, penalising our Precision and causing the system to lower Recall.
- **Future work**: Transform the retrieved word domains to domain-specific word senses to complete the task.

REFERENCES

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. Revising the Wordnet Domains Hierarchy: semantics, coverage and balancing. In *Proc. of the Workshop on Multilingual Linguistic Resources*, MLR2004, 2004.

James R. Curran. *From Distributional to Semantic Similarity*. PhD thesis, Univ. of Edinburgh, 2004.

William A. Gale, Kenneth W. Church, and David Yarowsky. One Sense Per Discourse. In *Proc. of the 5th DARPA Speech and Nat. Lang. Workshop*, 1992.

Rob Koeling, Diana McCarthy, and John Carroll. Domain-Specific Sense Distributions and Predominant Sense Acquisition. In *Proc. of the Conf. on Human Language Technology and Empirical Methods in NLP*, HLT/EMNLP-05, 2005.

Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics*, COLING/ACL-98, 1998.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. The role of domain information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4):359–373, 2002.

William Webber, Alistair Moffat, and Justin Zobel. A Similarity Measure for Indefinite Rankings. *ACM Trans. on Info. Systems*, 28(4):20:1–20:38, Nov. 2010.