Making sense of word senses:
An introduction to word-sense disambiguation and induction

Alfredo Maldonado
ADAPT Centre at Trinity College Dublin
alfredo.maldonado@adaptcentre.ie
@alfredomg
http://alfredomg.com

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Agenda

- What are word senses and why are they problematic?
- (Classic) WSD Formulation
- Some Interesting Approaches
  - Supervised
  - Unsupervised
- Beyond WSD/I – new approaches and interpretations of the problem

Sense and Sensibility

- What does *sense* mean?
- The Oxford English Dictionary says:
  - … any of the faculties of sight, hearing, smell, taste and touch…
  - Natural understanding or intelligence, esp. in relation to practical matter arising in everyday life… (→ common sense)
  - Written or spoken discourse that is sensible, coherent, or readily intelligible.
  - A direction, esp. one of two opposite directions.
  - The meaning of a more or less extended sequence of written or spoken words (as a sentence, passage, book, etc.)
  - The meaning of a written or spoken word, compound, or short phrase. Also: any of the various meanings of a word or short phrase; the meaning of a word in a particular collocation or context.
  - … (OED lists 26 senses for *sense*, many with 2-3 subsenses)

Word-senses

- The word *sense* is polysemous:
  - i.e. it has more than one sense
- Many, if not most words, in any given language will be polysemous
- 121 most common English nouns have on average 7.8 WordNet senses (Ng and Lee 1996)
- Polysemous words are problematic for NLP systems – without WSD:
  - MT systems would mistranslate many words
  - Search Engines would return pages relating to irrelevant meanings of words in search queries
  - ….

Word-sense disambiguation

- WSD usually not seen as an end in itself, but as a service to other NLP tasks
  - MT, syntactic parsing, semantic parsing, information retrieval, information extraction, knowledge acquisition, …
- Should it be a drop-in black box or should it be integrated (perhaps implicitly) within a larger NLP task?
- WSD is an AI-complete problem (Ide and Véronis 1998)
  - To solve WSD we need to have complete natural language understanding or common-sense reasoning
  - Context can give us a clue to the meaning of words (Weaver 1949; 1955)

Approaches to WSD

- Dictionary-based (aka knowledge-based)
  - Lesk (1986) Algorithm compares ambiguous word's dictionary definitions to ambiguous word's context
  - Use of selectional preferences to choose appropriate sense
- Supervised corpus-based
  - Usage of annotated corpora with word senses
  - Includes semi-supervised, bootstrapping methods
- Unsupervised methods
  - Word-sense discrimination/induction (Schütze 1998)
  - Word context clustering: each induced cluster represents an induced sense
- Hybrids
  - Any combination of the methods above
  - Translational equivalence (using multilingual parallel corpora)
  - Combination of WSD with other tasks. Entity Linking and WSD in Babelfy (Moro et al. 2014)

- WSD is traditionally formulated as a classification problem:
  - Given the instance of a word (in a sentence or paragraph), determine its sense from a given list of senses (from a dictionary or thesaurus)

$$\hat{s} = \arg\max_{s_i} P(s_i | context(w_j))$$

$$s_i \in S(w_j) = \{s_1, \ldots, s_n\}$$

- Each word has its own set of senses
- One classifier per ambiguous word

Some interesting approaches (Supervised)

- Naïve Bayes (Mooney 1996; Ng 1997; Leacock et al. 1998; Pedersen 1998; Bruce and Wiebe 1999)

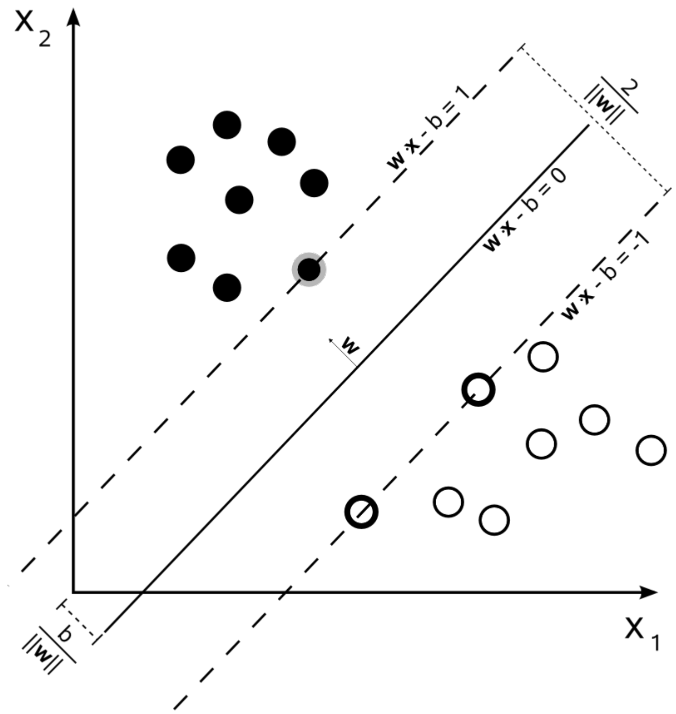
$$s_i \in S(w_j) = \{s_1, \ldots, s_n\}$$

$$\begin{aligned}
\hat{s} = \arg\max_{s_i} P(s_i | context(w_j)) &= \arg\max_{s_i} P(s_i | w_{j-l}, \ldots, w_{j+l}) \\
&= \arg\max_{s_i} \frac{P(s_i) P(w_{j-l}, \ldots, w_{j+l} | s_i)}{P(w_{j-l}, \ldots, w_{j-1}, w_{j+1}, \ldots, w_{j+l})} \\
&= \arg\max_{s_i} P(s_i) \prod_{k=j-l}^{j+l} P(w_k | s_i)
\end{aligned}$$

- For word window of size l
- Features can include POS, syntactic dependencies, etc.
- Assumption: features conditionally independent given the sense
- Beats "most frequent sense" baseline
- Performs well for classical supervised formulation of WSD

Some interesting approaches (Supervised)

- Support Vector Machines (SVM) (Escudero et al. 2000; Murata et al. 2001; Keok and Ng 2002)



Plot By Cyc - Own work, Public Domain,
https://commons.wikimedia.org/w/index.php?curid=3566688

- Learns the hyperplane that maximally separates senses
- Multiclass case by learning several binary SVMs (one sense against all others) – select sense class with best confidence
- Linear and polynomial kernels perform similarly
- Beats Naïve Bayes and nearly all other methods for classic supervised formulation of WSD

Some interesting approaches (Supervised)

- Neural Networks
  - Words connected as input nodes (Cottrell 1989)
  - Mapping nodes in hidden layers to dictionary definitions or WordNet concepts (Véronis and Ide 1990; Tsatsoranis et al 2007)
  - Perceptrons without hidden layers, using local and topical features (Towell and Voorhees 1998)
  - Deep Belief Networks (DBN) using a probabilistic generative model with multiple layers of hidden units (Wiriyathammabhum 2012)
  - sense2vec: Separates each sense of a word into separate context embeddings (Trask et al. 2015)
- Perform better than Naïve Bayes but not always better than SVMs (DBN beats SVMs)
- Require large amounts of training data

Word-Sense Induction (Unsupervised)

- No pre-defined list of word senses
- Task is to induce clusters and each cluster is interpreted to be a sense
  - Sense clusters are not readily interpretable
  - Attempts to make clusters human readable by automatically generating descriptive labels from co-occurring words (Kulkarni and Pedersen 2005)
  - WSI → WSD: By mapping clusters to tagged senses based on score maximisation (Munkres 1957; Purandare and Pedersen 2004)
- OR given two instances of an ambiguous word in context, determine whether the word is used in the same sense or a different sense

Some interesting approaches (Unsupervised)

- Word Space (Schütze 1998; Purandare and Pedersen 2004)
  - First-order co-occurrence context vectors (c1) & Second-order co-occurrence context vectors (c2)
    - Both represent instances of ambiguous words in context
    - Each dimension in c1 counts direct co-occurrences

      Money kept in the bank is safe

      c1(money) =

river	flow	money	kept	bank	safe
0	0	0	1	1	1

      c1(kept) =

river	flow	money	kept	bank	safe
0	0	1	0	1	1

    - All c1s in the corpus for a word can be aggregated (summed or averaged) to compute a **word vector** for that word

      w(bank) =

river	flow	money	kept	bank	safe
120	80	125	50	12	200

    - The c2 for a word is the sum of the word vectors co-occurring with it

      c2(bank) = w(money) + w(kept) + w(safe)

      =

river	flow	money	kept	bank	safe
14	22	685	382	50	545

  - c1 and c2 vectors can be optionally SVD-reduced, weighted by IDF, etc.

Some interesting approaches (Unsupervised)

- Word Space: c1s vs c2s ?
  - In general, c1s tend to perform better than c2s for WSI
  - c2s introduce a lot of noise
  - However, if dataset is really really small, then c2s can perform better than c1s.

Some interesting approaches (Unsupervised)

- Latent Semantic Analysis (Furnas et al 1988; Deerwester et al. 1990)
- Originated in Information Retrieval (Latent Semantic Indexing)
- But was adapted to represent lexical semantics for different tasks
- Didn’t take long before it was used for WSD/I (Levin et al. 2006)
- You have an index for your corpus, i.e. a matrix A in which each cell a_{ij} that counts how many times a word t_i occurs in a document d_j:

$$\mathrm{A} = \begin{array}{c} \\ \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_\mathbf{m} \end{array} \begin{array}{c} \begin{array}{ccc} \mathbf{d}_1 & \cdots & \mathbf{d}_\mathbf{n} \end{array} \\ \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \end{array}$$

- Each column is a document vector. Each row is a word vector.
- IR people use columns. Semantics people use rows.

Some interesting approaches (Unsupervised)

- In LSA, you factorise A using Singular Value Decomposition (SVD)

$$\hat{A} = U_k D_k V_k$$

- U_k columns are first k left-singular vectors of A
  - Used to project document vectors to reduced space
- V_k columns are first k right-singular vectors of A
  - Used to project word (term) vectors to reduced space

- SVD-reduced spaces capture higher order co-occurrence and are able to *handle* synonyms

Some interesting approaches (Unsupervised)

- You say that Word Space can be SVD-reduced
- LSA involves SVD
- Both involve vectors and matrices
- Are Word Space and LSA the same thing?
  - No. But they're very related. (Maldonado and Emms 2012; Maldonado 2015)
  - It is possible to convert objects from one system to the other via Linear Algebra trickery

Corpus-based WSD/I Results

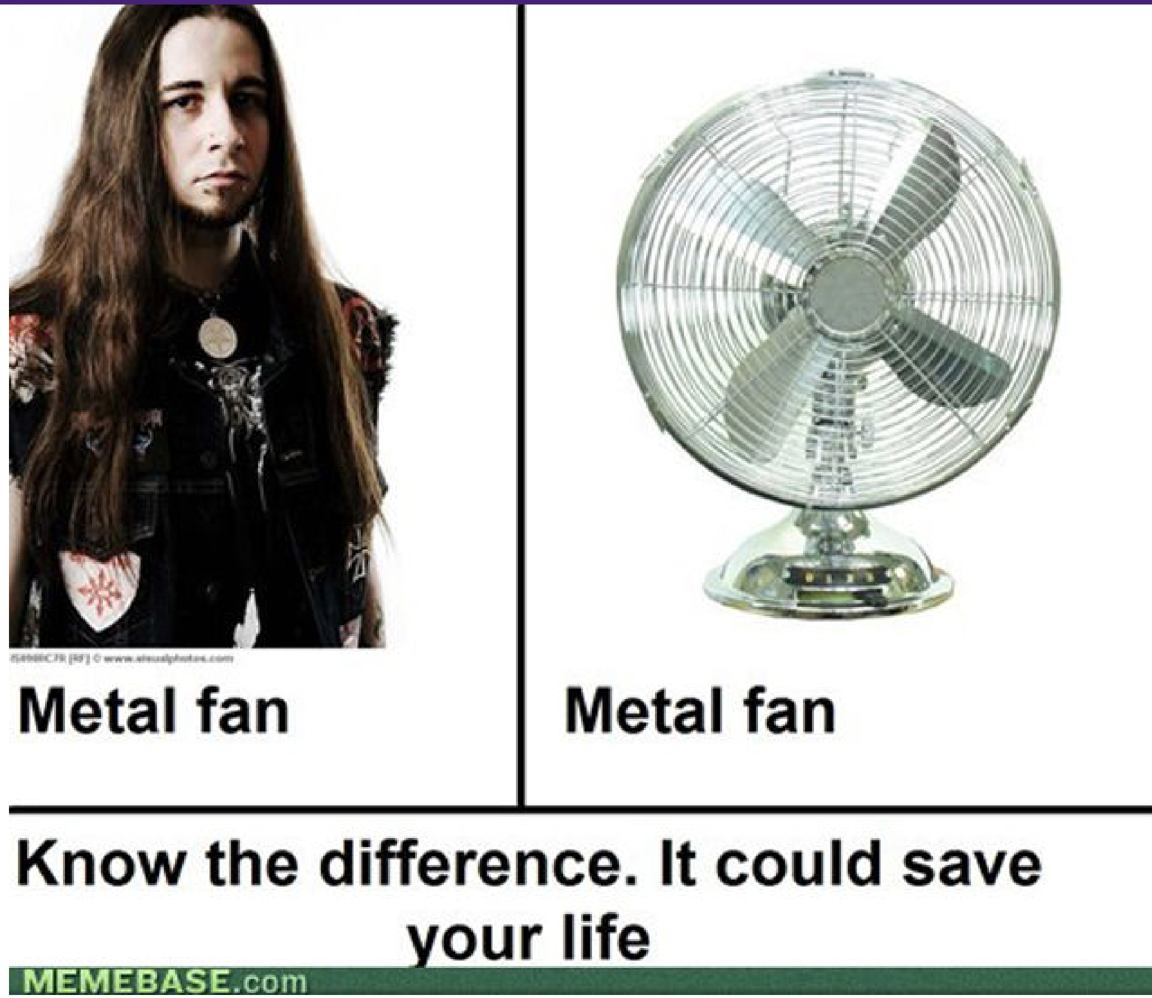
- Supervised approaches superior to dictionary-based approaches and unsupervised approaches
- Type and scope (local vs topical) of contextual feature heavily depends on type of word to disambiguate:
  - Nouns: wide context and local collocations
  - Verbs: syntactic features
  - Homographs much easier than polysemous words
- Given a predefined list of senses, state of the art methods perform very close to humans in traditional supervised formulation
- Word senses are subjective – different dictionaries will divide up the senses of the same word differently. What about domain-specific senses of a term? Word senses depend on the purpose of the task involving word senses (Kilgarriff 1997)
- Knowledge Acquisition Bottleneck in supervised WSD

  (Agirre & Edmonds 2007; Navigli 2009)

Beyond WSD/I

- Cross and Multilingual WSD
- All words WSD
  - Traditional approach: One classifier per word. Assumes word senses are independent.
  - Disambiguate each word depending on the sense of neighbouring words (WSD as a sequence labelling problem?)
  - Data sparseness problem: most words will appear only once in corpus, consequence of zipf's law
- Named Entity Disambiguation/Discrimination
- Babelfy – named-entity linking
- Research in WSD / SemEval competitions have spawned lots of semantic tasks:
  - Semantic Role Labelling
  - Sentiment Analysis
  - Textual Entailment
  - Cross-level semantic similarity
  - Semantic (Dependency) Parsing
  - ...

Questions?



alfredo.maldonado@adaptcentre.ie

@alfredomg alfredomg.com

References

Agirre, E., & Edmonds, P. (2007). Word Sense Disambiguation: Algorithms and Applications. (E. Agirre & P. Edmonds, Eds.). Springer.

Bruce, R. And Weibe. J. (1999). Decomposable modeling in natural language processing. Comput. Ling. 25, 2, 195–207.

Deerwester, S. et al. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407.

Escudero et al. (2000). On the portability and tuning of supervised word sense disambiguation. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC, Hong Kong, China). 172–180.

Furnas, G. W., et al. (1988). Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. In Proceedings of the 11th Anual International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 465–480, Grenoble.

Ide, N., & Véronis, J. (1998). Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. Computational Linguistics, 24(1), 1–40.

Keok, L and Ng, H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP, Philadelphia, PA). 41–48.

Kilgarriff, A. (1997). I don't believe in word senses. Computers and the Humanities, (31):91–113.

Kulkarni, A. and Pedersen, T. (2005). SenseClusters: Unsupervised Clustering and Labeling of Similar Contexts - Appears in the Proceedings of the Demonstration and Interactive Poster Session of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 105-108, June 26, 2005, Ann Arbour, MI.

Leacock et al. (1998). Using corpus statistics and WordNet relations for sense identification. Computat. Ling. 24, 1, 147–166.

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In Proceedings of the 1986 ACM SIGDOC Conference (pp. 24–26). Toronto.

Levin, E. et al. (2006). Evaluation of Utility of LSA for Word Sense Discrimination. In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, Companion Volume: Short Papers, number 1998, pages 77– 80, New York, NY.

References

Maldonado, A. and Emms, M. (2012). First-order and second-order context representations: geometrical considerations and performance in word-sense disambiguation and discrimination. In Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012). pp. 675-686. Liège.

Maldonado, A. (2015). Linear transformations of semantic spaces for word-sense discrimination and collocation compositionality grading. PhD Thesis. Trinity College Dublin.

Mooney, R.J. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP). 82–91.

Moro, A., Raganato, A., & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics, 2, 231–244.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. Journal of the Society of Industrial and Applied Mathematics, 5(1):32–38.

Murata, M. et al. (2001). Japanese word sense disambiguation using the simple Bayes and support vector machine methods. In Proceedings of the 2nd International Workshop on EvaluatingWord Sense Disambiguation Systems (Senseval-2, Toulouse, France). 135–138.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. ACM Computing Surveys, 41(2), 1–69.

Ng, H. T., & Beng Lee, H. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (pp. 40–47). Santa Cruz, Ca.

Ng, T. H. (1997). Getting serious about word sense disambiguation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics:Why, What, and How? (Washington D.C.). 1–7.

Pedersen, T. (1998). Learning probabilistic models of word sense disambiguation, Ph.D. dissertation. Southern Methodist University, Dallas, TX

Purandare, A. and Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In Proceedings of the Conference on Computational Natural Language Learning, pages 41–48, Boston, MA.

Schütze, H. (1998). Automatic word sense discrimination. Computational Linguistics, 24(1), 97–123.

Trask, A. et al. (2016). sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings. arXiv:1511.06388 [cs].

References

Weaver, W. (1955). Translation. In W. N. Locke & A. D. Booth (Eds.), Machine translation of languages: fourteen essays (pp. 15–23). Cambridge, MA: MIT Press.

Wiriyathammabhum, P. et al. (2012). Applying deep belief networks to word sense disambiguation, arXiv preprint arXiv:1207.0396