Computational Semantics Evaluation: The Origins of Senseval and Evolution of SemEval

Diana McCarthy
University of Cambridge (DTAL)
diana@dianamccarthy.co.uk

16th July 2018

Outline

Computational Semantics

- Ascribing meaning to linguistic units
- Units: words, phrases, sentences, discourse
- 'Meaning':
  - word senses (*match*)
  - categories (*bread* is-a food)
  - semantic relationships (*wheel* part-of *bus*), (*give birth* before$\rightarrow$ *die*) (*murder* cause$\rightarrow$ *death*),
  - logical form/thematic role (*I give the dog a bone*)
    `give(x,y,z) I(x) dog(y) bone(z) /`
    `agent=I, recipient = dog, theme = bone`
  - idiomatic usage (*he gave them a run for their money*)
- How do (should) we measure success?

Evaluation Goals and Considerations

- Why this task?
  - Is the task assumed useful for an application?
  - Does the task model human language?
- The objective:
  - To decide if the task itself is viable
  - To measure success
  - To compare systems
- What measurements should we use?
  accuracy, precision, recall, coverage, correlation ...
- What is the best we can expect?
- What is the worst we should expect?
- On what data?
  - Availability (licenses)
  - Bias

Outline

Word Sense Disambiguation (WSD)

Given a word in context, find the best-fitting "sense"

Residents say militants in a station wagon pulled up, doused the building in gasoline, and struck a match.

After the match, replace any remaining fluid deficit to prevent problems of chronic dehydration throughout the tournament.

Word Sense Disambiguation (WSD)

Given a word in context, find the best-fitting "sense"

Residents say militants in a station wagon pulled up, doused the building in gasoline, and struck a match.

match#n#1

After the match, replace any remaining fluid deficit to prevent problems of chronic dehydration throughout the tournament.

match#n#2

Senseval 1998

- International evaluation 'exercise' (not competition)
  followed from discussions at
  Tagging Text with Lexical Semantics (Light, 1997)
- "nearly as many [WSD] test suits as there are researchers"
  (Resnik and Yarowsky, 1997)
- Need for standardized approach 'gold standard' to level the playing field
  - data and sense inventories
  - measure how 'gold' is gold using agreement between humans
  - metrics: precision and recall
- inevitable biases, but aim to make comparisons possible

Senseval(s)

1998,2001,2004

- Methodology focusing on conditions (including deadlines)
- Multiple languages
- Inclusivity (avoided labelling as a 'competition')
- Sampling choices: All words vs lexical sample tasks
- Sense inventories:
  - Hector (unseen)
  - WordNet (widely available)
  - Distinctions from parallel data (Japanese Translation Task 2001)
- Hand-labelled blind and Inter-tagger(annotator) agreement to estimate upper-bound
- Random and most frequent baselines

Word Sense Disambiguation: The Reality

WSD seems appealing, but what about

- *line* – mark long rather than wide vs adjacent/queue vs text
- *bar* – piece of metal or wood vs counter(in pub) vs pub
- *child* - offspring vs young person vs . . .

> *You fall in love or give birth to a child, and suddenly you remember the miracle of existence. (UKWaC)*

Coarse-grained Senses

- Many believe we need a coarse-grained level for WSD applications (Ide and Wilks, 2006) (though see Stokoe (2005))
- But what is the right way to group senses?

Example *child* WordNet

WNs#	gloss
1	a young person
2	a human offspring
3	an immature childish person
4	a member of a clan or tribe

- For machine translation (MT) use parallel corpora if you know the target languages
- But what about other applications such as summarising, paraphrasing, question answering (QA), information retrieval?

Coarse-grained Senses

- Many believe we need a coarse-grained level for WSD applications (Ide and Wilks, 2006) (though see Stokoe (2005))
- But what is the right way to group senses?

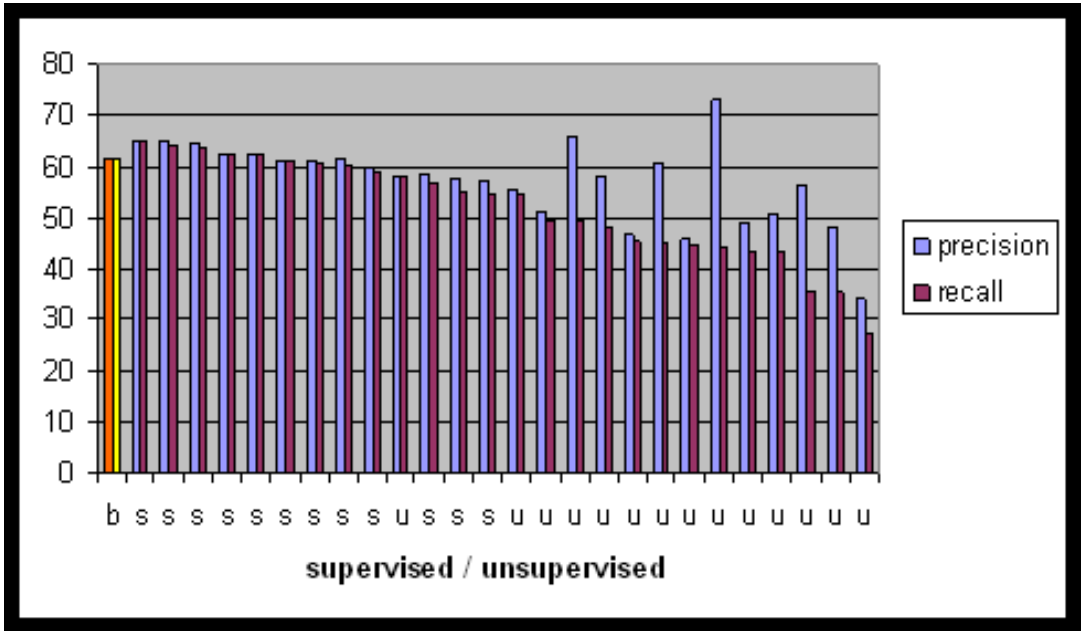
Example *child* WordNet SENSEVAL-2 groups

WNs#	gloss
1	a young person
2	a human offspring
3	an immature childish person
4	a member of a clan or tribe

- For machine translation (MT) use parallel corpora if you know the target languages
- But what about other applications such as summarising, paraphrasing, question answering (QA), information retrieval?

Senseval Findings

- Supervised systems outperform "unsupervised" but costly
- They need a large quantity of hand-labelled data
- Performance just better than first sense heuristic
  e.g. English all words Senseval-3 results



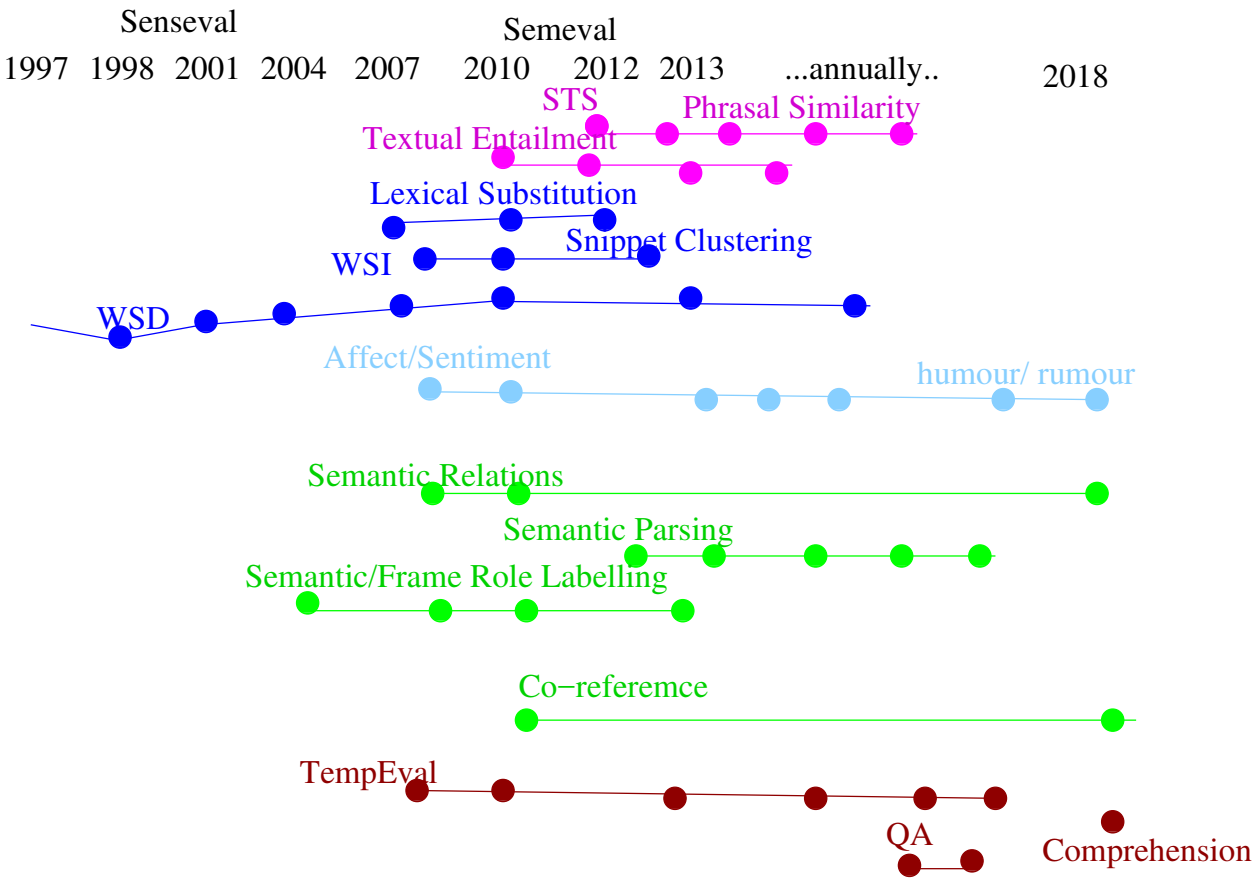
Senseval-3 Panel discussion

- plateau in performance, (reflecting skew of the data and bias)
- not learning "what works well on what"
- desire to demonstrate utility of WSD (issue for parsing too!)
- desire to encourage new ideas/tasks (diversity)
- NB also desire for inter-operability (2007 discussion)

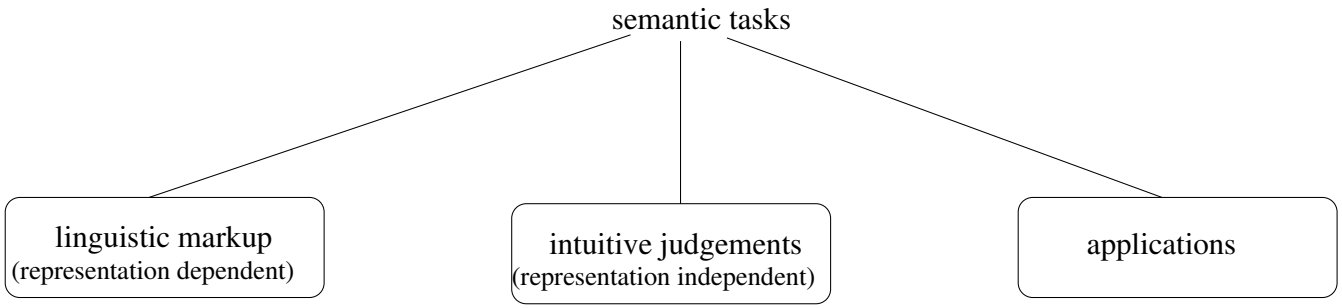
Outline

The Evolution of SemEval

A sample of tasks



A SemEval Task Categorization

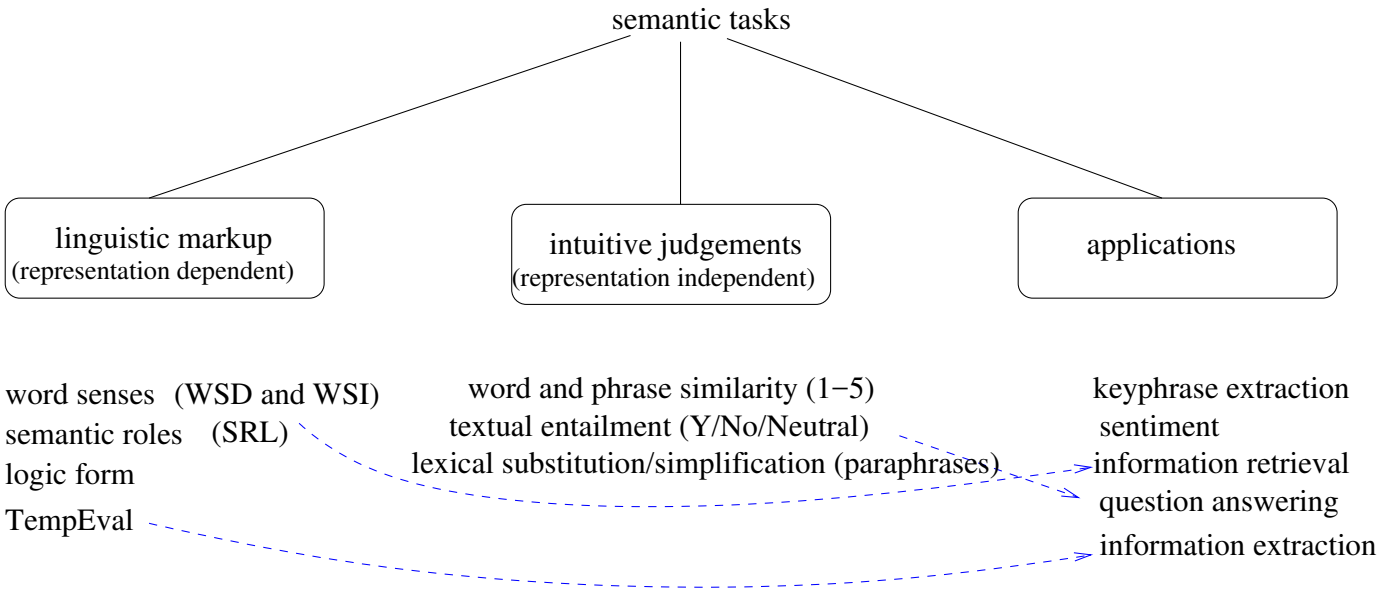


word senses (WSD and WSI)
semantic roles (SRL)
logic form
TempEval

word and phrase similarity (1−5)
textual entailment (Y/No/Neutral)
lexical substitution/simplification (paraphrases)

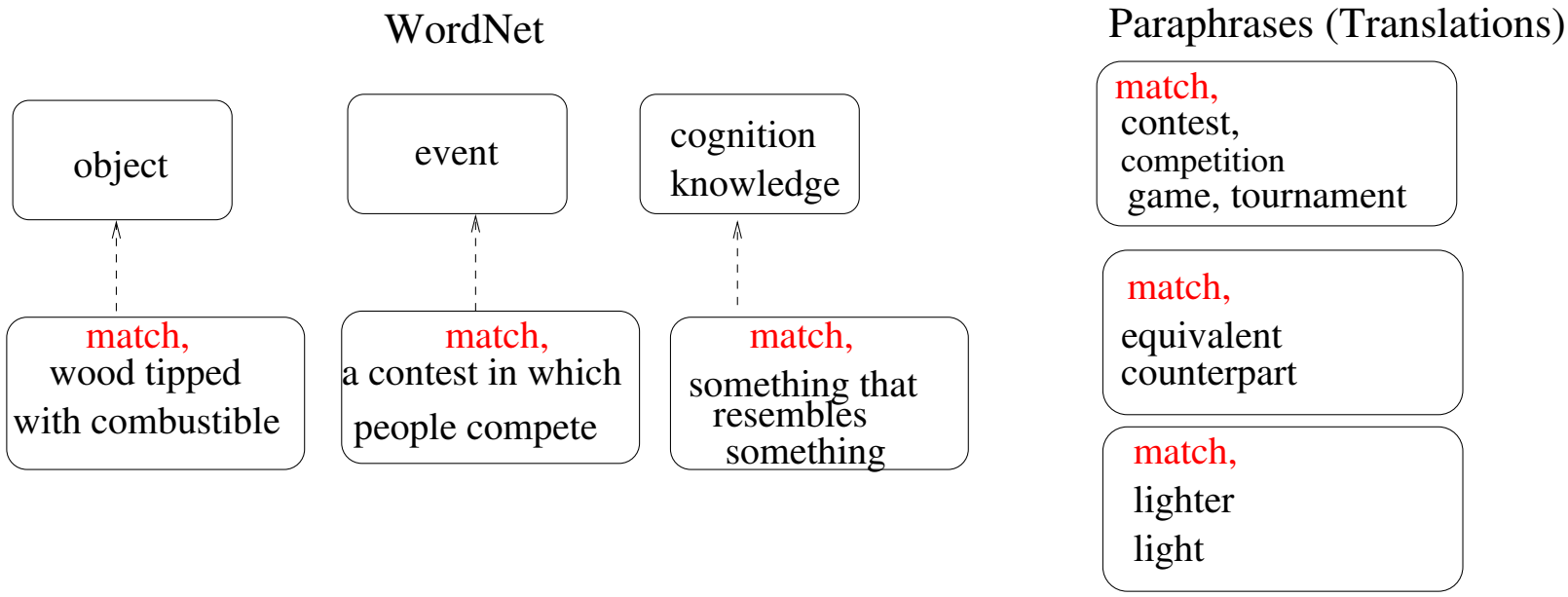
keyphrase extraction
sentiment
information retrieval
question answering
information extraction

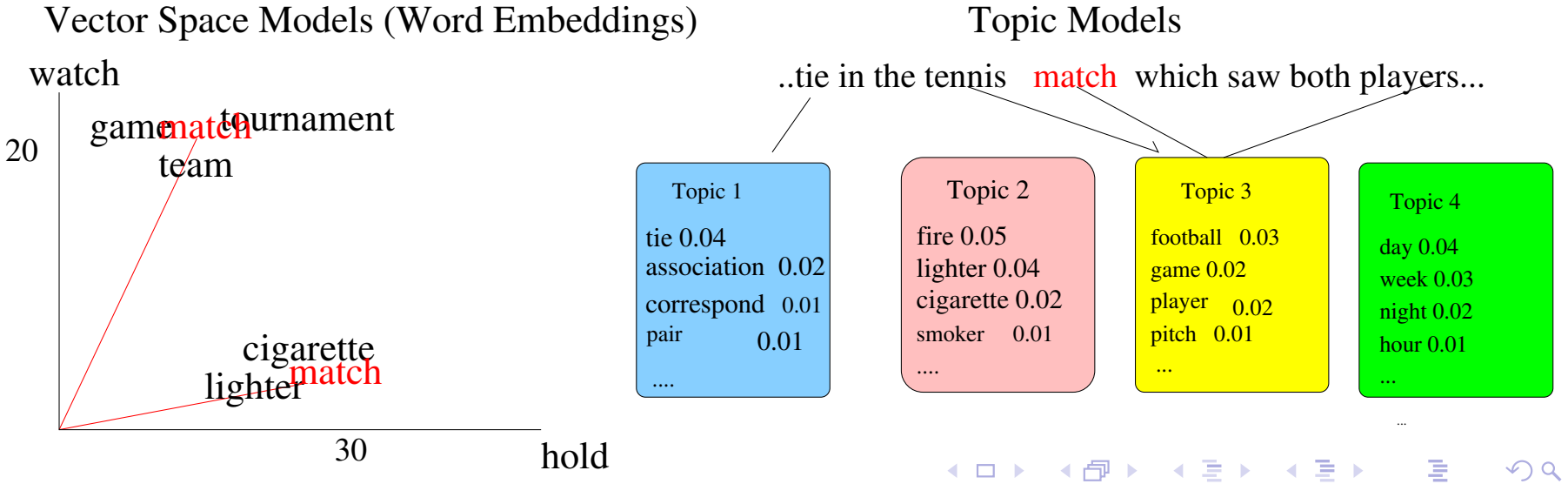
A SemEval Task Categorization



Word Sense Representation and WSI

(Schütze, 1998; Apidianaki et al., 2014; Lau et al., 2012; Brody and Lapata, 2009)





Word Sense Induction: How to Evaluate?

SemEval 2007, 2010 and 2013

- Compare clusters to those induced from traditional tagging

sentence	gold	system
S1	WN2	SYS1
S2	WN1	SYS1
S3	WN1	SYS1
S4	WN2	SYS2

- WSD task using best possible mapping
- Clustering metrics which compare gold with system clustering
- 2013 included a graded task which allows soft clusters and weights
- However just as for WSD , what is the right inventory?
- None of these tasks were truly representation independent as the gold-standard is based on WordNet distinctions

Word Sense Induction: How to evaluate?

WSI and WSD within an End-User Application (Navigli and Vannella, 2013)

- Clustering web snippets, for example *apple*:
  1. *Apple Inc., formerly Apple Computer, Inc., is...*
  2. *The science of apple growing is called pomology...*
  3. *Apple designs and creates iPod and iTunes...(Annotations Wikipedia senses)*
- Wikipedia disambiguation page as inventory (botany, companies, film and television, music etc...)
- Topic modelling trained using wikipedia best performance (Lau et al., 2013)!

Lexical Substitution (McCarthy and Navigli, 2007)

Find a replacement word for a target word in context

For example

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match , followed by a warm-up hit and perhaps a top-up with extra fluid before the match.

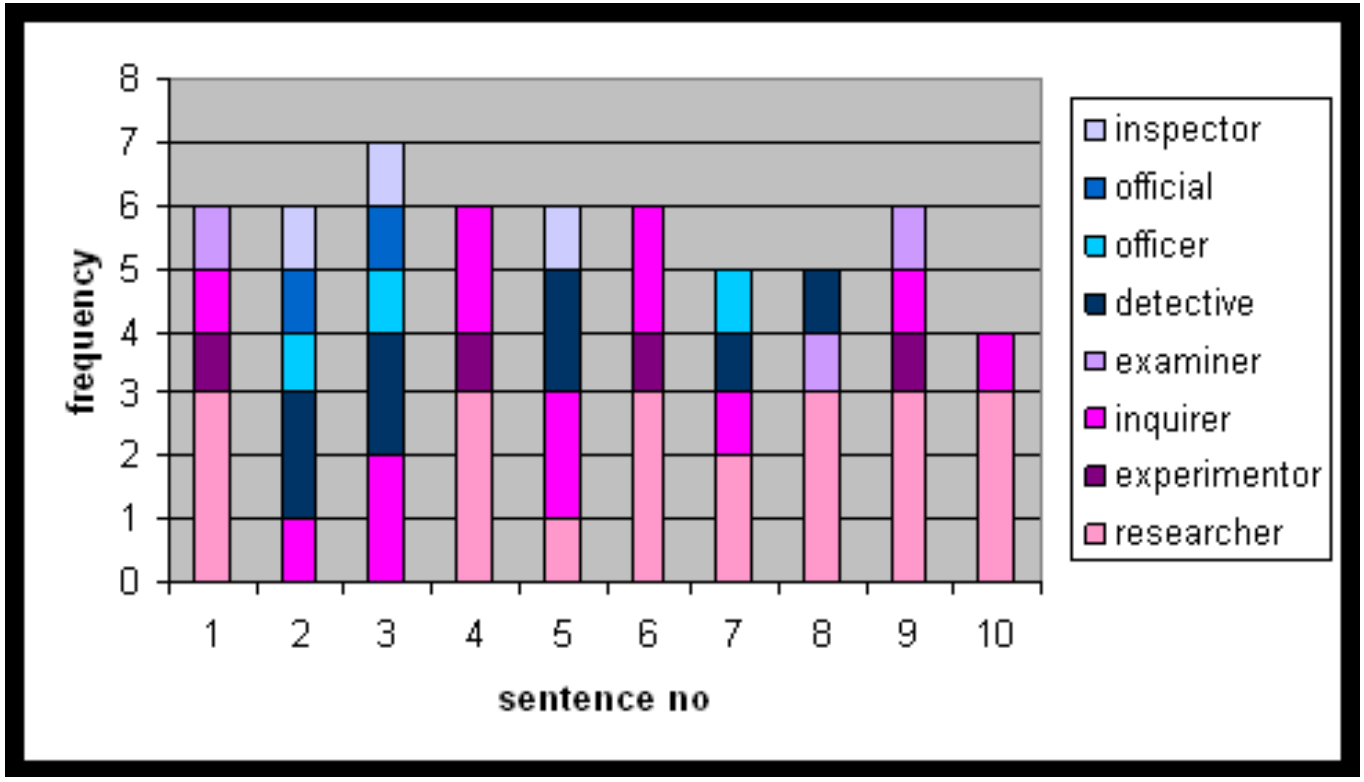
Lexical Substitution (McCarthy and Navigli, 2007)

Find a replacement word for a target word in context

For example

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match , followed by a warm-up hit and perhaps a top-up with extra fluid before the game.

Substitutes for *investigator* (noun)



Cross-Lingual Lexical Substitution (Mihalcea et al., 2010)

Example *solid.a*

- 1082: *We are confident, that by signing the treaty, the friendly relationship between the two countries has become solid.*
- 1083: *Glacius - Pour some water into your hand and cause it to freeze solid*
- 1087: *Huge areas that had been solid enough for camping a day earlier were now saturated with water.*

S	LEXSUB substitutes	CLLS translations
1082	dependable 1;strong 1;firm 1;cemented 1;genuine 1;stable 1;reliable 1;	fuerte 4;solido 4;resistente 1;
1083	hard 3;rigid 1;set 1;	solido 4;tempano 1;congelado 1;en estado solido 1;
1087	firm 2;hard 2;strong 1;dry 1;set 1;sound 1;	fuerte 2;resistente 1;macizo 1;consistente 1;solido 1;seguro 1;duro 1;firme 1;

Outline

Logical Form, Frames and Semantic Roles

Example: *Popeye said, 'I usually eat spinach'*

- Senseval-3 Logical Forms (Rus, 2004)
  Popeye:n(x1) say:v(e1,x1,e2) eat:v(e2,x1,x2) spinach:n(x2) usually:r(e2)
- SemEval 2007 FrameNet task (Baker et al., 2007).
  Annotate using semantic frames and frame elements from FrameNet (Ruppenhofer et al., 2010)

Lexical Unit	Frame	Frame Element (Roles) - Filler
say	Statement	Speaker - Popeye, Message - *I usually eat spinach*
eat	Ingestion	Ingestor - I, Ingestibles - Spinach

Abstract Meaning Representation Parsing and Generation (May and Priyadarshi, 2017; May, 2016)

The London emergency services said that altogether 11 people had been sent to hospital for treatment due to minor wounds.

```
(s / say-01
      :ARG0 (s2 / service
            :mod (e / emergency)
            :location (c / city :wiki ''London''
                  :name (n / name :op1 ''London'')))
      :ARG1 (s3 / send-01
            :ARG1 (p / person :quant 11)
            :ARG2 (h / hospital)
            :mod (a / altogether)
            :purpose (t / treat-03
                  :ARG1 p
                  :ARG2 (w / wound-01
                        :ARG1 p
                        :mod (m / minor)))))
```

other related tasks e.g. 2019 Cross-Lingual Semantic Parsing with UCCA (Universal Conceptual Cognitive Annotation)

Phrasal Similarity (Korkontzelos et al., 2013)

Two subtasks:

1. Semantic Similarity of word and 2-word phrase (Y/N)

   demeanor – *non verbal behaviour*
   aubergine – *psychotic disorder*

2. detecting non-composition

Never go back to a lit firework - it may go off in your face.
The musical backing is not in your face like some of today's recordings
. . .

Semantic Textual Similarity (Agirre et al., 2012, 2016a)

- Judgements on Scale (1-5) between two text fragments (sentences)
- Relevant to applications e.g. Question Answering, Summarisation, Machine Translation
- Data from a variety of domains including:
  - plagiarism corpus
  - Q/A Question-question and answer/answer similarity
  - image descriptions
- Monolingual (2012) and then Cross-Lingual (2016)

Semantic Textual Similarity (Agirre et al., 2016a)

Guidelines:

- 5 Two sentences are completely equivalent
- 4 Most equivalent (unimportant difference in detail)
- 3 Roughly equivalent but important info missing or differs
- 2 Two sentences are not equivalent but share some details
- 1 Two sentences are not equivalent, but are on the same topic
- 0 Two sentences are completely dissimilar.

1. *They flew out of the nest in groups.*
2. *They flew into the nest together*

Semantic Textual Similarity (Agirre et al., 2016a)

Guidelines:

- 5 Two sentences are completely equivalent
- 4 Most equivalent (unimportant difference in detail)
- 3 Roughly equivalent but important info missing or differs
- 2 Two sentences are not equivalent but share some details
- 1 Two sentences are not equivalent, but are on the same topic
- 0 Two sentences are completely dissimilar.

1. *They flew out of the nest in groups.*
2. *They flew into the nest together*

Interpretable Semantic Textual Similarity (Agirre et al., 2015, 2016b)

On what grounds is something similar?

- Alignments
- Relation (similar, more specific/general, equivalent)
- Similarity score

1. *12 killed in bus accident in Pakistan*
2. *10 killed in road accident in NW Pakistan*

Interpretable Semantic Textual Similarity (Agirre et al., 2015, 2016b)

On what grounds is something similar?

- Alignments
- Relation (similar, more specific/general, equivalent)
- Similarity score

1. *12 killed in bus accident in Pakistan*
2. *10 killed in road accident in NW Pakistan*

```
[12] <=> [10] : (SIMILAR 4)
[killed] <=> [killed] : (EQUIVALENT 5)
[in bus accident] <=> [in road accident] : (MORE-SPECIFIC 4)
[in Pakistan] <=> [in NW Pakistan] : (MORE-GENERAL 4)
```

Textual Entailment

- Tasks run alongside and intersecting with SemEval
- Important for natural language inference (summarization, question-answering, Machine Translation evaluation . . .)
- Unidirectional decision from premise sentence to hypothesis
- Data construction: given premise sentence people asked to produce three hypotheses (Entailment, Contradiction, Neutral)

For example, premise:
A man reads the paper in a bar with green lighting.

1. *The man is inside.* E
2. *The man is reading the sportspage.* N
3. *The man is climbing a mountain.* C

SICK: Sentences Involving Compositional Knowledge (Marelli et al., 2014)

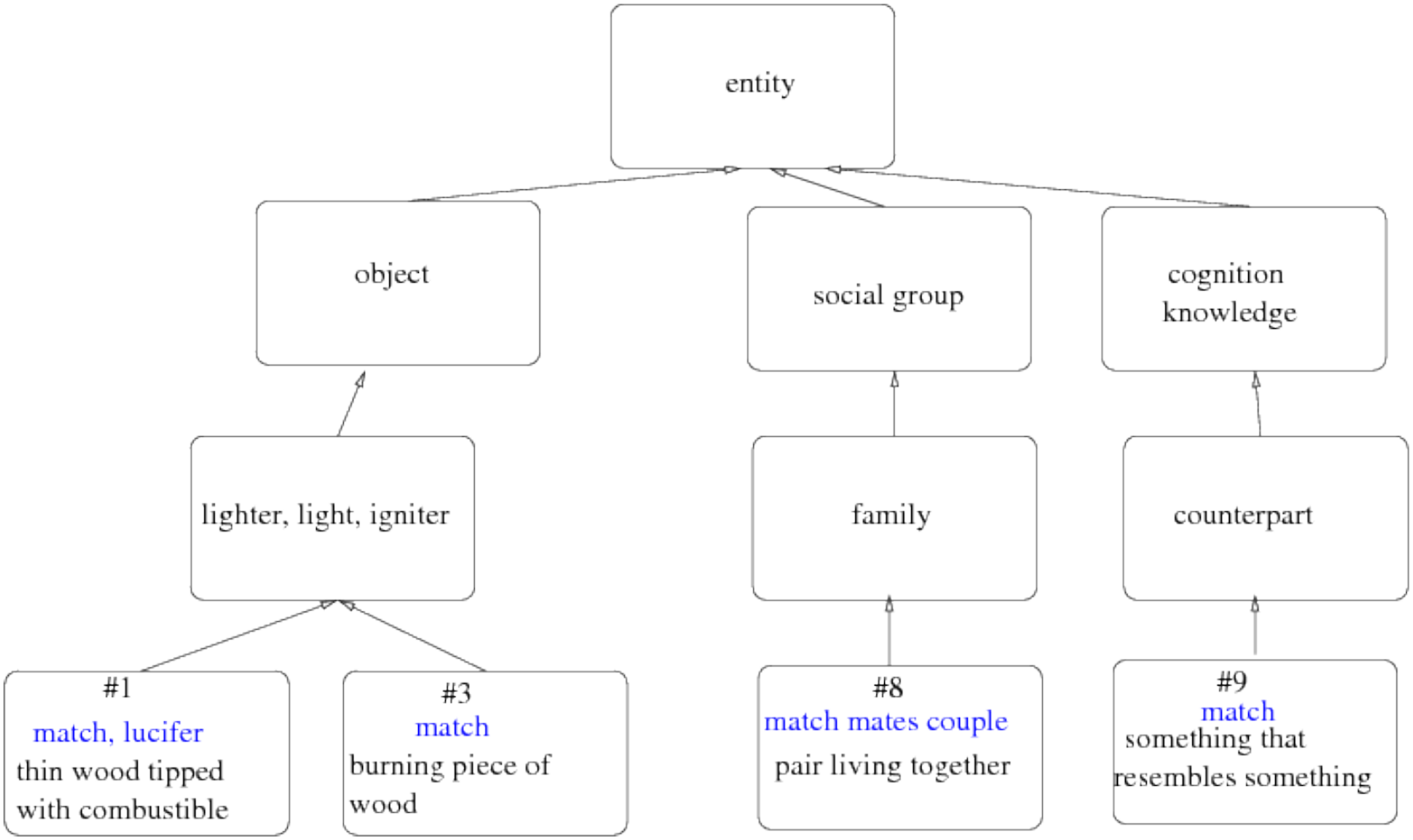
- Two tasks, operating on sentence pairs:
  - Semantic Relatedness 1-5
  - Entailment (Entailment, Contradiction and Neutral)
- Focus on composition, rather than multiwords, named entities and encyclopedic information:
  forked out – *buy*, *EU* – *European Union*, *Paris* vs France
- Focus on issues e.g. negation active/passive which are not frequent in STS and textual entailment datasets
- Finding: systems exploit ad-hoc features e.g. negation and antonyms to detect contradiction

Evaluating Computationals Semantics: Representation

- Semantics is covert, so how should we annotate for evaluation?
  - standard representations e.g. WordNet senses, FrameNet Frames and Roles, AMR mark up?
  - representation independent e.g. similarity, Y/N, paraphrases
    - allow comparison of different approaches with less bias☺
    - minimal guidelines for annotators ☺
    - but need careful analysis of data to see where the faults lie ☹
- and note . . .

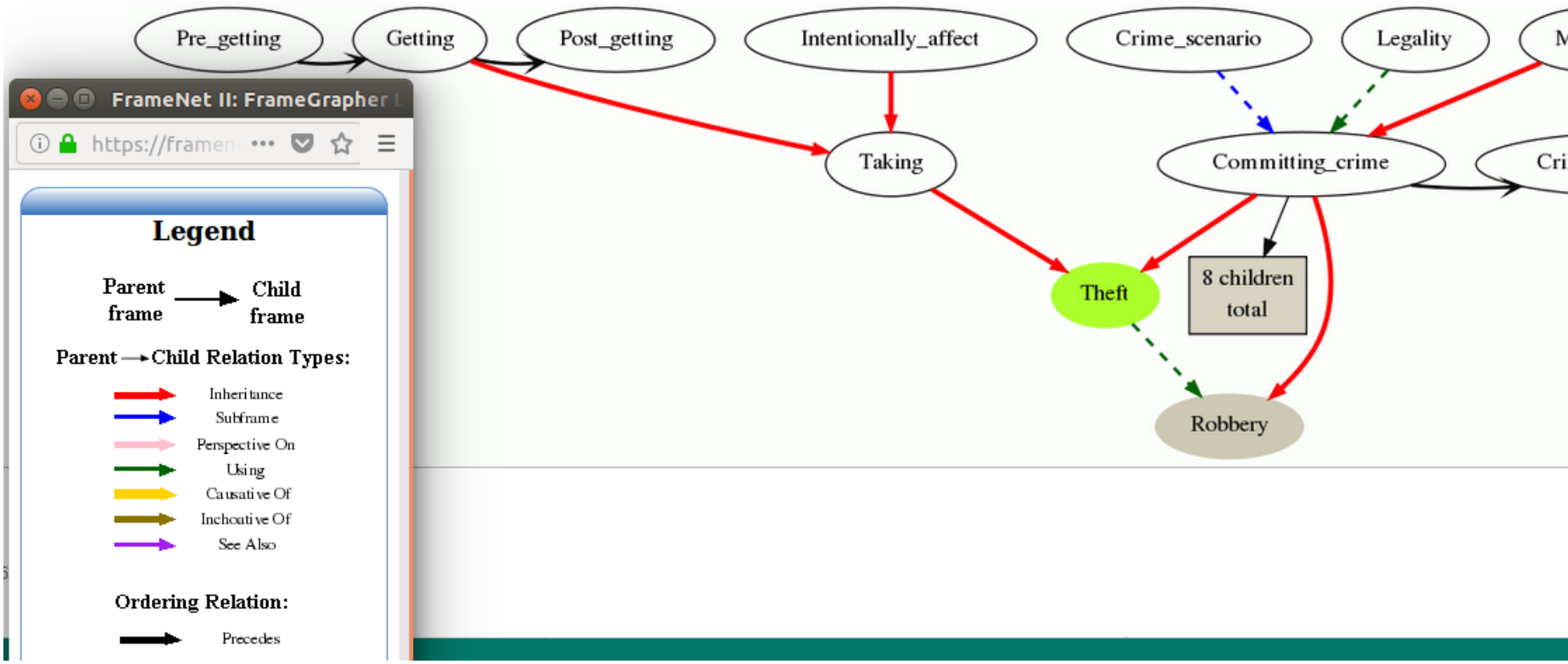
Representation may be help downstream applications

WordNet provides Semantic Relationships



Representation may help downstream applications

FrameNet may help with inferences



Pros and Cons

- Representation dependence:
  - bias to that theory/inventory ☹
  - costly to produce annotations ☹
  - annotations may allow hooks to other semantic resources ☺
- Representation independence
  - fairer comparison of very different approaches ☺
  - intuitive and easier to elicit (crowd source (Biemann, 2013)) ☺
  - interpretation may require careful scrutiny ☹
- Human agreement depends on task
- Possibilities for different types of annotation on the same data (without over-engineering for a dataset)

And Finally...

- Senseval and SemEval offer a wide variety of datasets available for many tasks and in many languages
- Useful for replicability (NB tasks are easier without deadlines)
- Evaluation to compare and learn (not just about headline scores)
- Watch for bias in datasets! Variety helps

And Finally...

- Senseval and SemEval offer a wide variety of datasets available for many tasks and in many languages
- Useful for replicability (NB tasks are easier without deadlines)
- Evaluation to compare and learn (not just about headline scores)
- Watch for bias in datasets! Variety helps
- Thank you for listening!

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016a). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Agirre, E., Cer, D., Diab, M., and Gonzalez, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In

*Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*.

Agirre, E., Gonzalez-Agirre, A., Lopez-Gazpio, I., Maritxalar, M., Rigau, G., and Uria, L. (2016b). Semeval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 512–524, San Diego, California. Association for Computational Linguistics.

Apidianaki, M., Verzeni, E., and McCarthy, D. (2014). Semantic clustering of pivot paraphrases. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4270–4275, Reykjavik, Iceland. European

Language Resources Association (ELRA). ACL Anthology Identifier: L14-1401.

Baker, C., Ellsworth, M., and Erk, K. (2007). Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104. Association for Computational Linguistics.

Biemann, C. (2013). Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Brody, S. and Lapata, M. (2009). Bayesian word sense induction. In *Proceedings of the EACL*, pages 103–111, Athens, Greece.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1).

Ide, N. and Wilks, Y. (2006). Making sense about sense. In Agirre, E. and Edmonds, P., editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL Workshop on multiword expressions: identifying and exploiting Underlying Properties*, pages 12–19.

Korkontzelos, I., Zesch, T., Zanzotto, F. M., and Biemann, C. (2013). Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Lau, J. H., Cook, P., and Baldwin, T. (2013). unimelb: Topic

modelling-based word sense induction for web snippet clustering. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 217–221, Atlanta, Georgia, USA. Association for Computational Linguistics.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France. Association for Computational Linguistics.

Light, M., editor (1997). *Proceedings of the* SIGLEX *Workshop on Tagging Text with Lexical Semantics: Why, What and How?*

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences

through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

May, J. (2016). Semeval-2016 task 8: Meaning representation parsing. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1063–1073, San Diego, California. Association for Computational Linguistics.

May, J. and Priyadarshi, J. (2017). Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545, Vancouver, Canada. Association for Computational Linguistics.

McCarthy, D. and Navigli, R. (2007). SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th*

International Workshop on Semantic Evaluations (SemEval-2007), pages 48–53, Prague, Czech Republic.

Mihalcea, R., Sinha, R., and McCarthy, D. (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

Navigli, R. and Vannella, D. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, Georgia, USA. Association for Computational Linguistics.

Resnik, P. and Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of*

the SIGLEX *Workshop on Tagging Text with Lexical Semantics: Why, What and How?*, pages 79–86, Washington, DC.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2010). FrameNet II: Extended theory and practice. Technical report, International Computer Science Institute, Berkeley. http://framenet.icsi.berkeley.edu/.

Rus, V. (2004). A first evaluation of logic form identification systems. In Mihalcea, R. and Edmonds, P., editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 37–40, Barcelona, Spain. Association for Computational Linguistics.

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

Stokoe, C. (2005). Differentiating homonymy and polysemy in information retrieval. In *Proceedings of the joint conference on*

Human Language Technology and Empirical methods in Natural Language Processing, pages 403–410, Vancouver, B.C., Canada.