# Textual Entailment
# Part 3: Knowledge Resources and Knowledge Acquisition


Sebastian Pado
Institut für Computerlinguistik
Universität Heidelberg, Germany

Rui Wang
Language Technology
DFKI, Saarbrücken, Germany


AAAI 2013, Bellevue, WA

Thanks to Ido Dagan for permission to use slide material

# This part of the tutorial

1. Overview: Types of Inference Knowledge
2. Use Case 1: Acquiring Asymmetrical Similarity
3. Use Case 2: Truth Status in Context

# Part 1: Types of Inference Knowledge



# Inference Rules

- TE assesses if H can be inferred from T
  - Requires linguistic knowledge, world knowledge
- Sentence-level entailment is always *decomposed* into atomic (subsentential) inference steps
  - Corresponding to *compositional* meaning construction
- Valid atomic inference steps can be represented as **inference rules** *a* → *b*
  - a, b almost arbitrary linguistic representations
  - Various linguistic levels (lexical, syntactic, phrasal, …)

# Application of Inference Rules

- Resources with inference rules are used in virtually every single Textual Entailment system:
  - Transformation-based approaches:
    Inference rules motivate proof steps
  - Classification-based approaches:
    Inference rules inform similarity features
- What types of inference knowledge is helpful?
  - Clark et al. (2006): analysis of knowledge types
  - Mirkin et al. (2009): ablation tests for various knowledge resources on entailment



# The Challenge for Knowledge

- Textual Entailment requires its inference rules to have both *high precision* and *high recall*
  – Low precision: rules do more harm than good
  – Low recall: rules are irrelevant

- Complementary behavior of resources:
  – Manually constructed resources often lack recall
  – Automatically constructed resources often lack precision

# Normalization Knowledge

- Named Entities, Abbreviations, Acronyms, etc.
  - Sources: Machine-readable dictionaries
  - Status: Relatively unproblematic

Mr. Clinton ⇔
Bill Clinton ⇔
President Clinton

US ⇔
U.S. ⇔
United States



# Lexical Knowledge

- **Nominal** relations: Synonymy, Hyponymy
  – Sources: WordNet, Distributional Thesauri
  – Status: most widely used type of knowledge, still recall problems
  **Use case 1**
- **Verbal** relations: Causation, Presupposition
  – Sources: WordNet, VerbOcean
  – Status: also widely used, but both recall and precision problems

Peter owns a kitchen **table** ⇒
Peter owns an **object**

Peter **buys** a kitchen table ⇒
Peter **owns** a kitchen table

# Syntactic Knowledge

- Structural variation (relative clauses, genitives, active/passive, etc.)
  - Sources: syntactic rule bases
  - Status: often used, but limited recall

Peter**, who** sleeps soundly, … ⇒
Peter sleeps soundly

Peter **broke** the vase. ⇒
The vase **was broken** by Peter.



# Paraphrase Knowledge

- Inferences that cannot be captured at word level
  - Variety of phenomena
  - Range from simple to very difficult
- Sources: Corpora (both monolingual and parallel)
- Status: Very difficult to balance precision and recall

X **buys** Y from Z ⇒
Z **sells** Y to X

X **gave** me **a hand** ⇒
X **helped** me

X was **a Yorkshireman by birth** ⇒
Y was **born in Yorkshire**

# World knowledge

- **Factual** Knowledge
  - Sources: Gazetteers, Wikipedia
- "**Core theories**"
  [Clark et al. 2006]
  - Sources: mostly hand-coded

T: Paris is **in France** ⇒
H: Paris is **in Europe**

T: Easter 2011 was **on April 24** ⇒
H: Easter 2011 was **between April 20 and 30**

- Status: Superficial treatment in most TE systems
  - Interesting direction: Unstructured vs. structured data – compare IBM Watson (Kalyanpur et al. 2012)



# Sentential Context

- Sentential context influences inference
- Variety of factors
  - Monotonicity
  - Clause Truth Status
    - **Use case 2**
  - Presupposition
- Status: Current research

T: Peter sees **a** poodle ⇒
H: Peter sees a dog

T: Peter sees **no** poodles ⇏
H: Peter sees no dogs

T: Peter **managed** to come ⇒
H: Peter came

T: Peter **promised** to come **⇒?**
H: Peter came

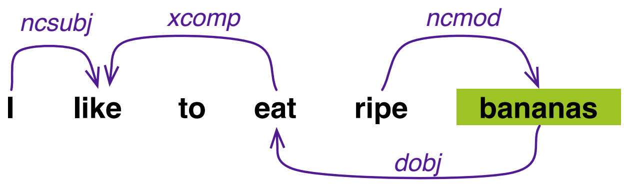
# Use Case 1: Asymmetrical Similarity (Kotlerman et al. 2010)

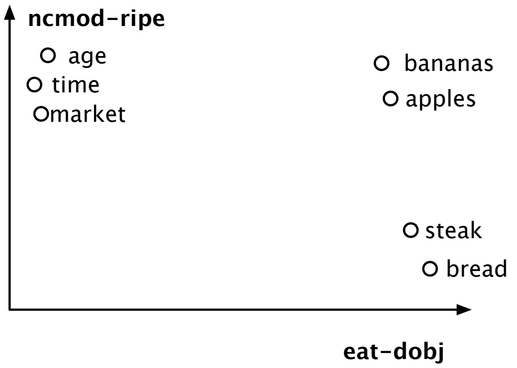


# Distributional Semantics

- Goal: Learn lexical inference rules *a* ⇒ *b* from corpora
- Distributional Semantics: “You shall know a word by the company it keeps” [Firth, 1957]
- An unsupervised way to model word meaning:
  - Observe in which contexts a word occurs
  - Represent words as vectors in high-dimensional space
  - Vector similarity correlates with semantic similarity
- Applied to many tasks in language processing [Turney & Pantel 2010]

# Distributional Similarity






# Standard Similarity Measures

- Cosine: Angle between vectors

$$cos(\vec{u}, \vec{v}) = \frac{\sum_i u_i \cdot v_i}{\sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2}}$$

- Lin's similarity: Pointwise mutual information of shared features

$$PMI(u, f) = \log \frac{P(u, f)}{P(u)P(f)}$$

$$lin(\vec{u}, \vec{v}) \frac{\sum_{i:u_i>0, v_i>0}[PMI(u, f_i) + PMI(v, f_i)]}{\sum_{i:u_i>0} PMI(u, f_i) + \sum_{i:v_i>0} PMI(v, f_i)}$$

# Acquiring Entailment Rules

- Standard approach: For each target word, find the highest-similarity neighbors
  - Synonyms (and other close semantic relations): Lexical entailment rules [Lin 1998]
  - Generalization from words to dependency paths: Paraphrase rules [Pantel and Lin 2001]



# Asymmetry of Inference Rules

- Standard similarity measures are symmetrical…
- …Inference rules are asymmetrical!

"Peter has a pet dog"

$\Rightarrow$ $\not\Leftarrow$

"Peter has a pet poodle"

***bank*** $\Rightarrow$ ***company***

***company*** $\not\Rightarrow$ ***bank***

# Symmetric Similarity - Results

- Most similar words for *food*:

| | | | |
|---|---|---|---|
| meat | clothing | water | sugar |
| beverage | foodstuff | coffee | material |
| goods | textile | meal | chemical |
| medicine | fruit | tobacco | equipment |

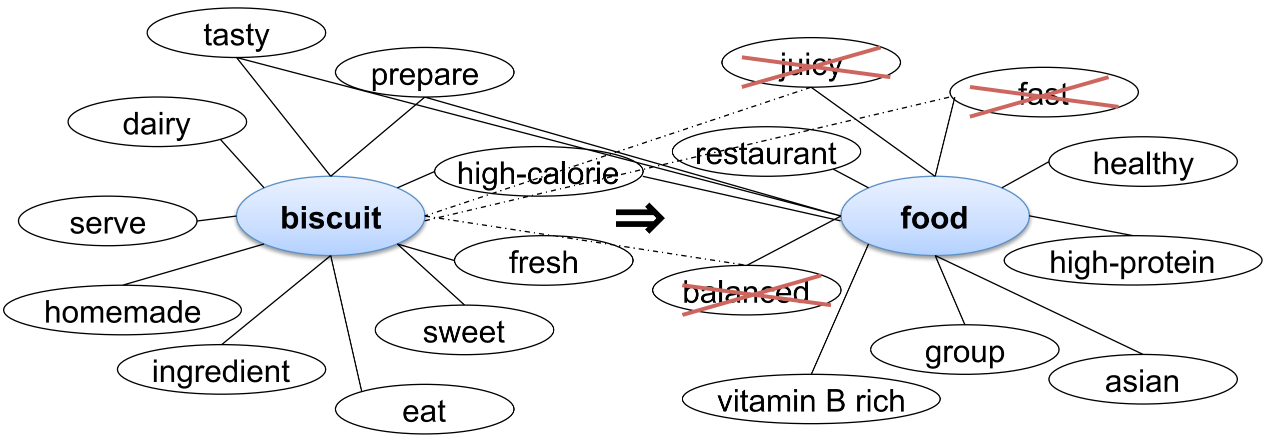
- Evaluation of resources for entailment (Mirkin et al. 2009)

| *Resource* | *Precision* | *Recall* |
|---|---|---|
| WordNet | 55% | 20% |
| Wikipedia | 45% | 7% |
| Dist.sim.(Lin) | 28% | 43% |



# Distributional Inclusion

- If u ⇒ v, then the characteristic contexts of u are expected to be characteristic for v, but not vice versa [Weeds et al., 2004]



# Average Precision

- Average Precision: Measure from Information Retrieval to assess search engine output (ranked list)
- Goals:
  - retrieve many relevant documents
  - retrieve few irrelevant documents
  - retrieve relevant docs early in list

| Retrieved | Relevant |
|---|---|
| Doc 1 | Doc 1 |
| Doc 2 | Doc 2 |
| Doc 3 | Doc 3 |
| Doc 4 | Doc 4 |
| Doc 5 | … |
| … | Doc 8 |
| ~~Doc 9~~ | Doc 10 |
| Doc 10 | … |
| … | Doc 299 |
| ~~Doc 300~~ | Doc 301 |
| … | … |

$$AP = \frac{\sum_i Prec(d_1, \ldots, d_i) \cdot rel(d_i)}{\sum_i rel(d_i)}$$

(where *rel* is 1 if doc is relevant)



# Balanced Average Precision

- Average Precision can applied to vectors to measure **feature inclusion**:
  - Retrieved, Relevant ⇒ u,v
  - Documents ⇒ Features
- u ⇒v if top features of v are shared by u and u has few other top features
- Modifications: *rel'* is graded relevance based on rank; balance with Lin similarity to alleviate sparse *v* vectors

| ***u*** ⇒ | ***v*** |
|---|---|
| Feature 1 | Feature 1 |
| Feature 2 | Feature 2 |
| Feature 3 | Feature 3 |
| Feature 4 | Feature 4 |
| Feature 5 | … |
| … | Feature 8 |
| ~~Feature 9~~ | Feature 10 |
| Feature 10 | … |
| … | Feature 299 |
| ~~Feature 300~~ | Feature 301 |
| … | … |

$$balAPinc(\vec{u}, \vec{v}) = \sqrt{lin(\vec{u}, \vec{v}) \cdot \frac{\sum_i Prec(f_1^u, \ldots, f_i^u) \cdot rel'(f_i^u)}{|\vec{u}|}}$$

# Directional similarity - results

- The most similar words to *food*:

| | | | |
|---|---|---|---|
| foodstuff | ration | blanket | margarine |
| food product | drinking water | soup | dessert |
| food company | wheat flour | biscuit | cookie |
| noodle | grocery | sweetener | sauce |
| canned food | beverage | meat | ingredient |
| feed | snack | agribusiness | meal |
| salad dressing | dairy product | diet | vegetable |
| bread | hamburger | medicine | vegetable oil |

- For more evaluation, see Kotlerman et al. (2010)



# Use Case 2: Truth Status in Context (Lotan et al. 2013)

# Motivation

- Reminder: Context influences entailment patterns
  - Complex phenomenon
- Subproblem: *Truth status* of *clauses* in context
  - Case 1: Clause is true (positively entailed) **[+]**
  - Case 2: Clause is false (negatively entailed) **[-]**
  - Case 3: Clause truth is unknown **[?]**

T: Peter managed to sleep. ⇒
H: Peter slept.

T: Peter failed to sleep. ⇒
H: Peter did not sleep.

T: Peter promised to sleep.⇒**?**
H: Peter slept.



# Determining Clause Truth

- Clause Truth is determined primarily by three factors:
  - Embedding words/phrases
  - Modifiers
  - Specific structures (presuppositions)
- Each factors can be associated with a *signature*
  - Description of its influence on clause truth

# Signatures

1. Negation: **[+] → [-], [-] →[+]**
2. Modality markers (many adverbs, modal verbs):
   **[+] → [?], [-] →[?]**
3. Factive embeddings (knowledge/perception/emotion):
   **[+] → [+], [-] →[+]**
4. Presuppositions (relative clauses, definite NPs): **[+]**
   [Kiparsky and Kiparsky 1970]
5. Implicative embeddings: various patterns
   [Karttunen 1971, 2012]
   – have the time: **[+] → [+], [-] →[-]**
   – make sure: **[+] → [+], [-] →[?]**

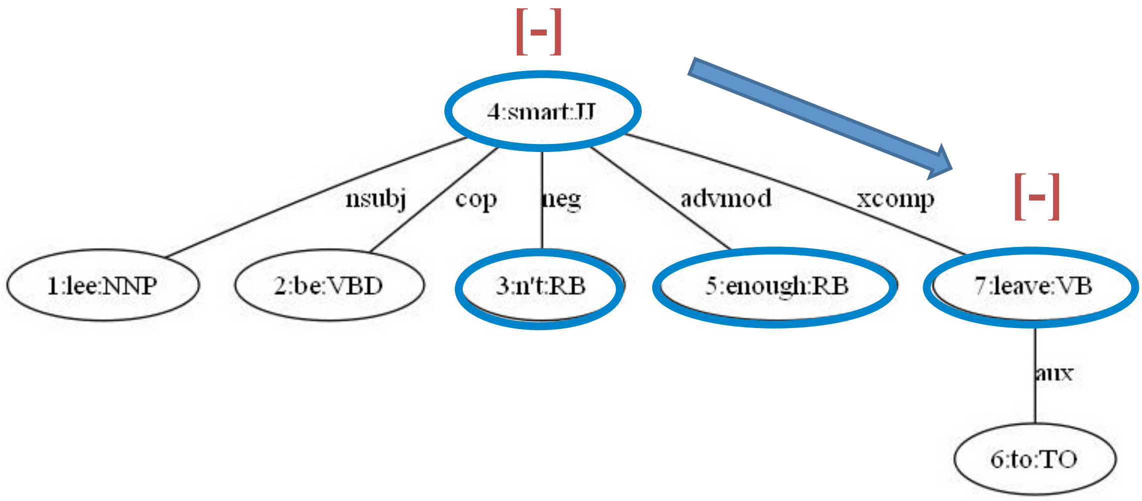


# The "TruthTeller" system

- Lexicon of modifiers and embedding words/phrases
  – About 2000 entries with signatures
  – Constructed semi-automatically
- Recursive algorithm inspired by Natural Logic
  [Lakoff 1970, MacCartney & Manning 2009]
  – Determine truth status of all clauses in a sentence

- http://u.cs.biu.ac.il/~nlp/downloads/TruthTeller/

# TruthTeller Example




Lee wasn't smart enough to leave



# Evaluation

- Evaluation against manual truth status labels
- Most frequent class baseline: 77% accuracy ( **[+]** )
- Total accuracy: 89%

| Truth Status | Recall | Precision | Occurrences |
|---|---|---|---|
| **[+]** | 87.3% | 98% | 120 |
| **[-]** | 74% | 83% | 50 |
| **[?]** | 91.4% | 70% | 48 |

- No evaluation integrated in RTE system yet

# Take-Home Messages

- Knowledge plays a central role in deciding TE
  - Can be represented uniformly with entailment rules
  - Multiple layers of linguistic and world knowledge
- Manual resources (coverage issue) vs. automatically acquired resources (accuracy issue)
- Use Cases:
  - Better automatic acquisition with asymmetrical similarity
  - More precise context modeling with truth status



# Reference List


- Clark, P., Murray, W. R., Thompson, J., Harrison, P., Hobbs, J. R., Fellbaum, C. (2007). On the Role of Lexical and World Knowledge in RTE-3. Proceedings of the ACL Workshop on Textual Entailment and Paraphrasing, 54–59.
- J.R. Firth (1957). Papers in Linguistics 1934–1951. London: Oxford University Press.
- Geffet, M. and Dagan, I. (2004). Feature Vector Quality and Distributional Similarity. Proceedings of COLING, 247-254.

# Reference List


- Kalyanpur, A., Boguraev, BK., Patwardhan, S., Murdock, JW. , Lally, A., Welty, C., Prager, JM. , Coppola, B., Fokoue-Nkoutche, A., Zhang, L. (2012). Structured data and inference in DeepQA. IBM Journal of Research and Development, vol. 56 3.4: 10–1.
- Karttunen, L. (1971). Implicative Verbs. Language (47), 340-58.
- Karttunen. L. (2012). Simple and Phrasal Implicatives. Proceedings of *SEM, 124-131.
- Kiparsky, P. and Kiparsky, C. (1970). Fact. In M. Bierwisch and K.E. Heidolph (eds), Progress in Linguistics, 143-73.




# Reference List


- Kotlerman, L., and Dagan, I., and Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. Natural Language Engineering 16(4), 359-389.
- Lakoff, G. (1970). Linguistics and natural logic. Synthese 22, 151-271.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. Proceedings of ACL/COLING, 768–774.
- Lin, D. and Pantel, P. (2002). Discovery of Inference Rules for Question Answering. Natural Language Engineering 7(4), 343–360.

- A. Lotan, A. Stern, and I. Dagan (2013). TruthTeller: Annotating Predicate Truth. Proceedings of NAACL, 752-757.
- MacCartney, W., and Manning, C. (2009). An extended model of natural logic. Proceedings of IWCS, 140-156.
- Mirkin, S., Dagan, I., and Shnarch, E. (2009). Evaluating the inferential utility of lexical-semantic resources. In Proceedings of EACL, 558–566.
- Turney, P. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. JAIR 37(1):141–188
- Weeds, J., Weir, D., McCarthy, D. (2004). Characterizing Measures of Distributional Similarity. Proceedings of COLING, 1015-1021.