Lexical Semantics

Semantics

- Semantics is the relationship between **signifiers** and their inherent meaning.
  - Signifiers include words, phrases, signs and symbols.
- What constitutes meaning?
  - **Denotation**
    - The de jure meaning of what's presented, text-literal explicit meaning.
  - **Connotation**
    - The de facto meaning of what's presented, socially-understood implicit meaning.
- Lexical Semantics
  - The study of semantics over **lexical units**, atomic pieces of a language, and their meaning relates to the language's **syntax**, the structure of the language.
  - Together, lexical units form a collection or catalogue called the **lexicon** of the language.

Meaning

- Realization of "The World"
  - What objects exist in the world?
  - What are those objects like?
  - What events have happened?
  - How do they relate?
    - The quick brown fox jumped over the lazy dog.
- Understanding
  - Inference and implication
  - Belief modeling
- Meaning As Action / Situated Meaning
  - Associating world experience to understanding.
    - Images, words, actions, procedures

Lexical Semantics

- The meaning of lexical units
    - Sense, reference
    - What is a *dog*?
    - Is a hound dog a dog?
    - Is a hotdog?
- Grammatical meaning
    - What do we know about *dog*?
        - The dog wagged its tail.
        - The dog is lazy.
        - The fox jumped over the dog.

Semantic Features

We can break apart a word's into multiple distinct features.

Man = { pos: noun, gender: male, species: human, age-category: adult }

Boy = { pos: noun, gender: male, species: human, age-category: child }

These act as a basis for identifying synonyms and antonyms.

Lexical Semantics

Lexical semantics is a field of linguistic semantics, and it focuses on the study of how and what the words of a language denote.

Two aspects of lexical semantics research:

Static: classification and decomposition of word meanings

Dynamic: study of word meanings in sentences

Example

"Steven P. Jobs is one of the company's co-founders and currently serves as its Chief Executive Officer. Mr. Jobs also has been a director of the Walt Disney company since May 2006." (Apple Inc. SCHEDULE 14A, 2010)

Task 1 Word sense disambiguation

find out exact meaning of each word, e.g., "Jobs" is a last name, not an "occupation" or "piece of work"

Task 2 Co-reference resolution

figure out which entity a noun word/phrase or pronoun refers to, e.g., strings of the same color refer to the same person

Word sense disambiguation is the process to determine which sense of a word is used in a given context

The noun “company” has 9 senses (first 8 from tagged texts)

1. (807) company -- an institution created to conduct business; "he only invests in large well-established companies"; "he started the company in his garage"

2. (64) company, troupe -- organization of performers and associated personnel (especially theatrical); "the traveling company all stayed at the same hotel"

3. (55) company, companionship, fellowship, society -- the state of being with someone; "he missed their company"; "he enjoyed the society of his friends"

4. (54) company -- small military unit; usually two or three platoons

5. (13) party, company -- a band of people associated temporarily in some activity; "they organized a party to search for food"; "the company of cooks walked into the kitchen"

6. (12) company -- a social gathering of guests or companions

7. (6) caller, company -- a social or business visitor

8. (1) company -- a unit of firefighters and equipment; "a hook-and-ladder company"

9. ship's company, company -- crew of a ship including the officers

The verb “company” has 1 sense (no senses from tagged texts)

1. company, companion, accompany, keep company -- be a companion to somebody

Motivating application: Machine Translation

I like her company since it offers great benefits.

“an institution created to conduct business”

“the state of being with someone”

我喜欢她的公司因为福利好。

我喜欢和她在一起因为有利可图。

I like her company benefits.

I like her because it is profitable.

Question answering, e.g., “Do you like her company?”

Current WSD Methods

As a critical computational linguistic task WSD was first studied in machine translation in the 1940s. Dozens of approaches and systems have been developed since then.

Knowledge based methods use dictionaries and thesauri, and context knowledge is extracted from glosses.

Supervised methods. Syntactic and semantic features are extracted from a sense-annotated training corpus to create a classifier.

Semi-supervised methods use a small annotated corpus as seed data in a bootstrapping process.

Unsupervised methods acquire contextual information from unannotated text, and senses can be induced using similarity measures.

Is WSD hard?

Knowledge is critical to WSD and very hard to acquire:

1. Coverage: 150,000 words, 6.18 senses/word on average

2. Evolving lexicon: ~2,500 new words/per year in English

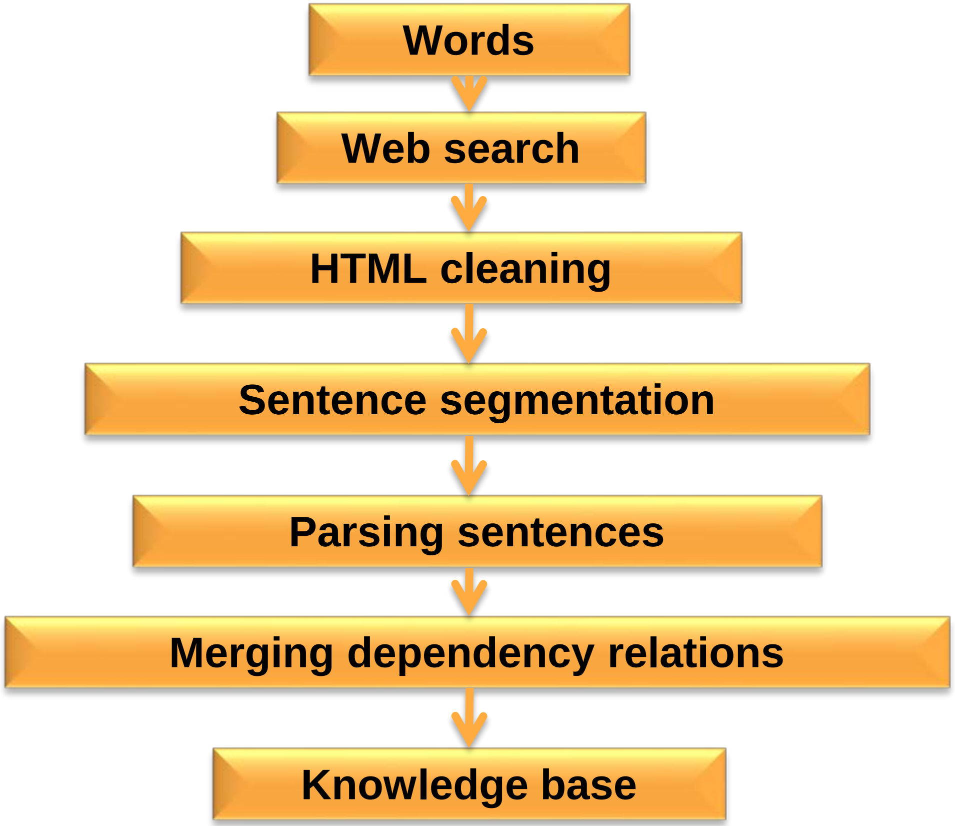
A practical WSD system needs:

Automatically acquirable WSD-capable knowledge of comprehensive coverage and constantly updated as the lexicon of a language evolves.

Machine-readable lexical knowledge base, e.g., WordNet

Unannotated text

WSD Knowledge acquisition process



Dependency Parsing

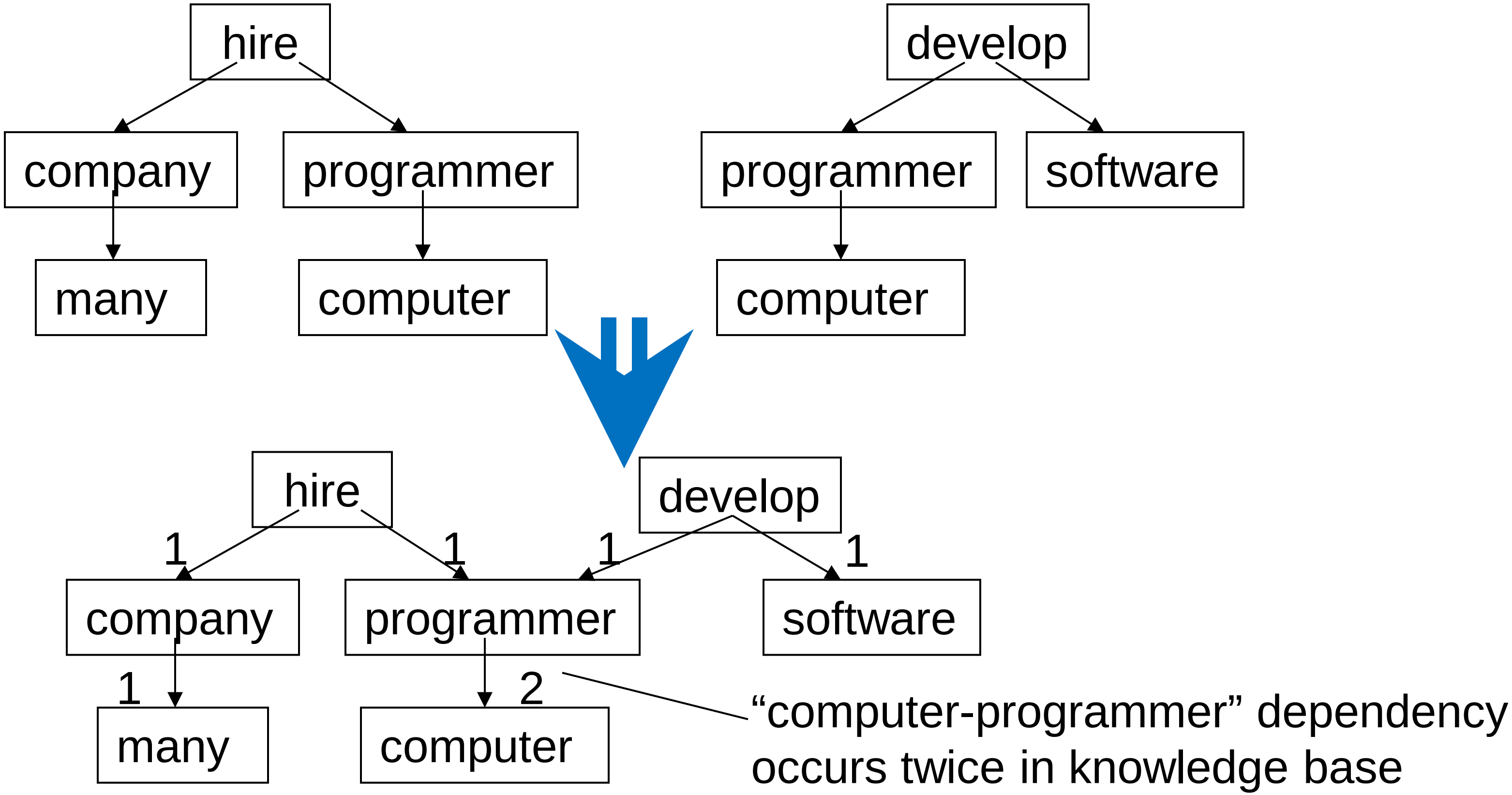
"Many companies hire computer programmers."

hire

company   programmer

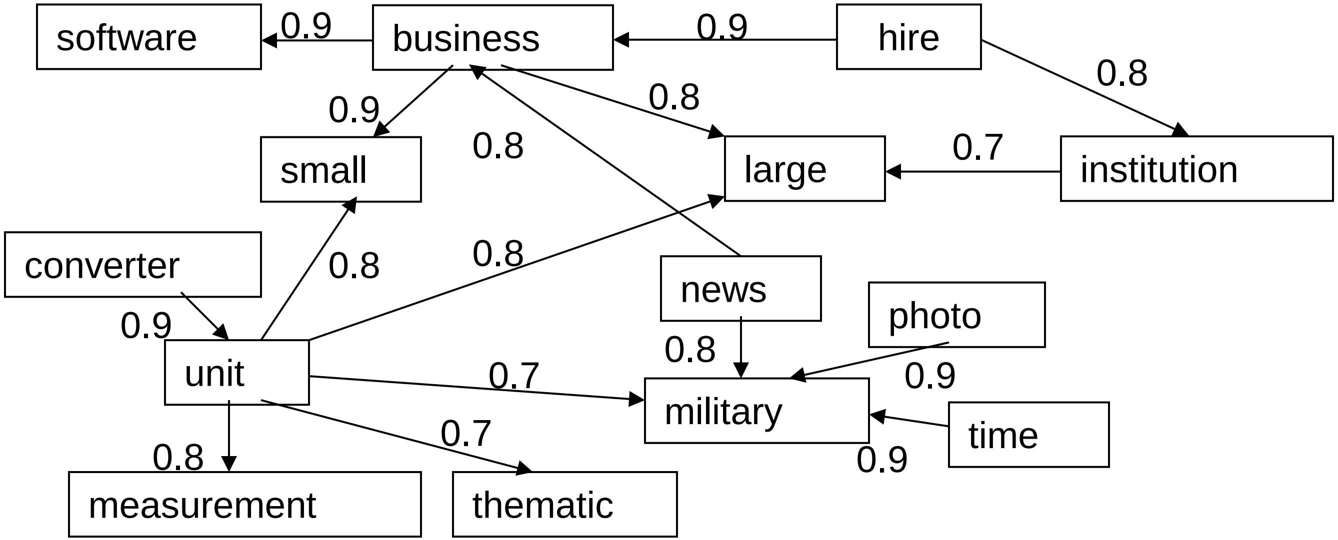
many   computer

Merging dependencies

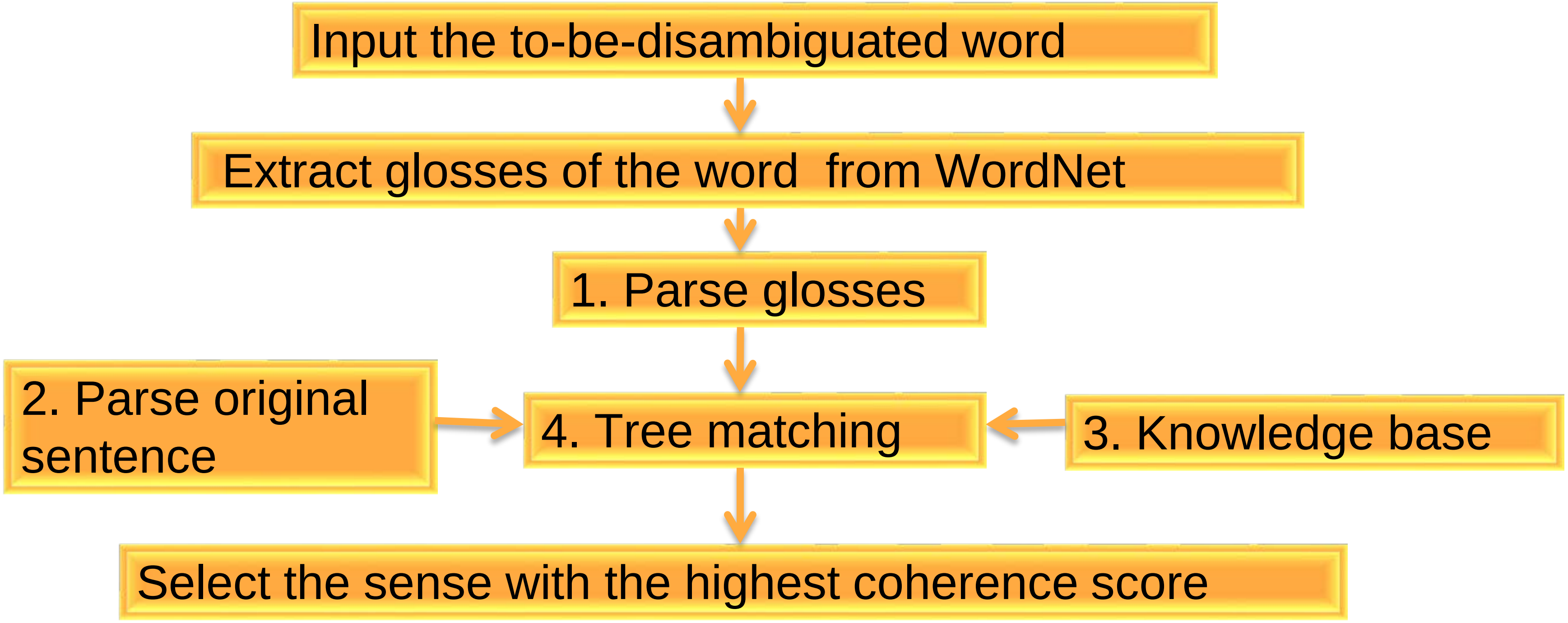


Normalized dependency knowledge

Using statistical significance test(e.g., Pearson's χ^2 test, Fisher's test), absolute frequency of a connection is normalized to a value Є [0,1] which denotes the semantic relevance of two words.



WSD process


Input the to-be-disambiguated word
Extract glosses of the word from WordNet
1. Parse glosses
2. Parse original sentence
4. Tree matching
3. Knowledge base
Select the sense with the highest coherence score

An example to disambiguate “company”

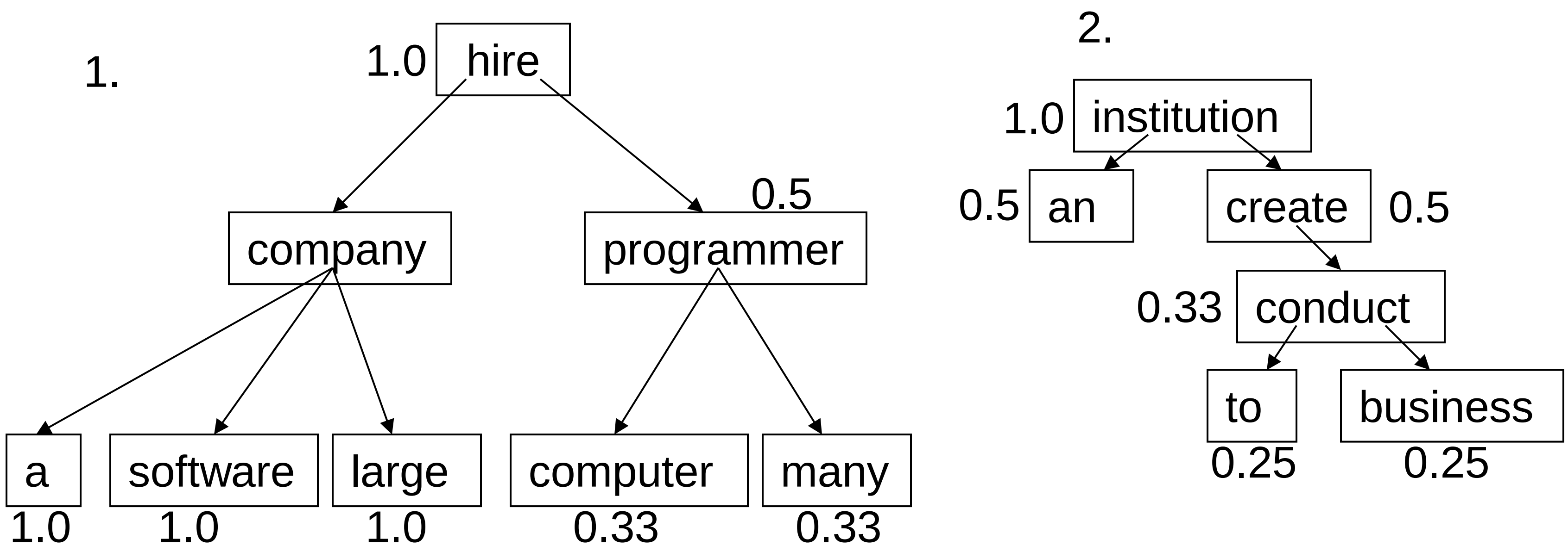
“A large software company hires many computer programmers.”

“company” has 9 senses as a noun in WordNet 2.1. Let's pick the following two glosses to go through our WSD process.

an institution created to conduct business

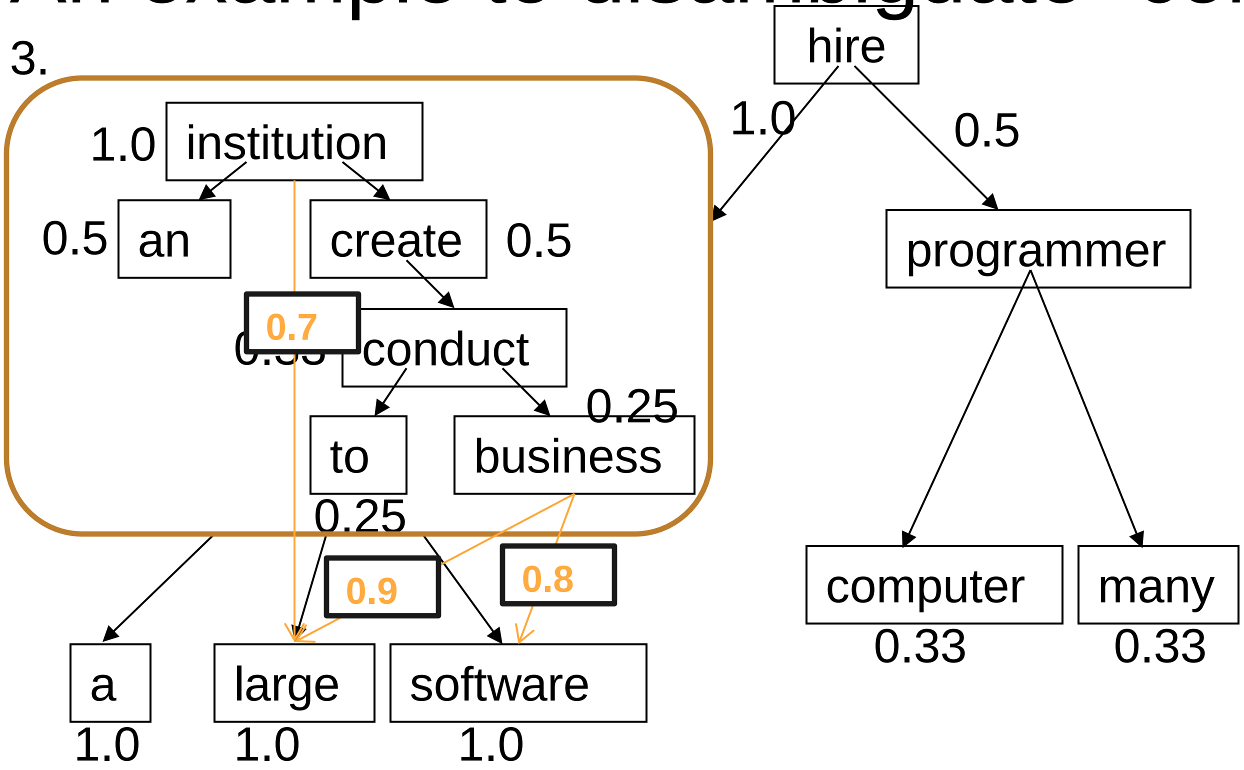
small military unit

An example to disambiguate “company”



An example to disambiguate “company”

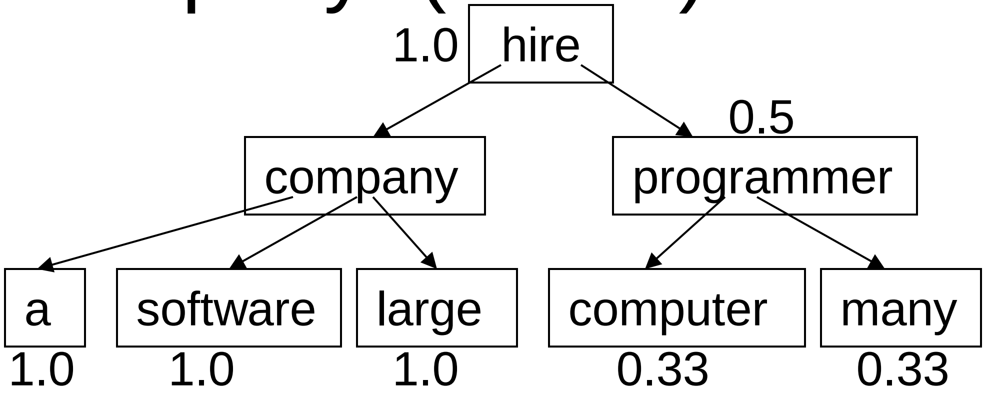
3.



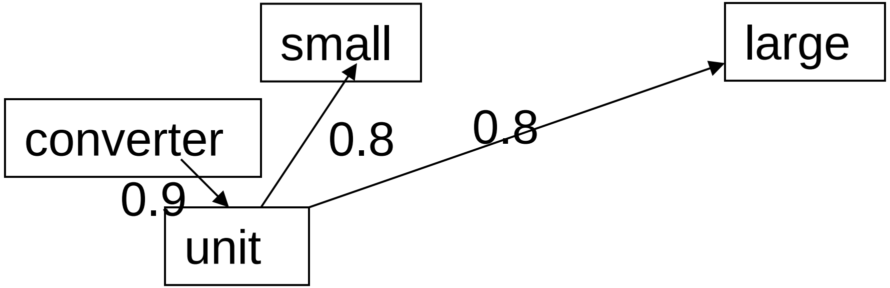
4. Semantic coherence score:

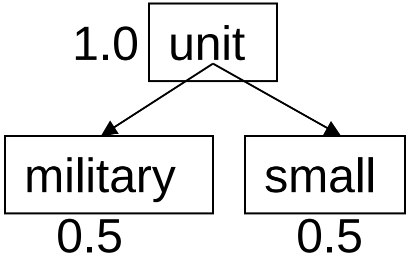
$$1.0 \times 1.0 \times 0.7 + 1.0 \times 0.25 \times 0.8 + 1.0 \times 0.25 \times 0.9 = 1.125$$

An example to disambiguate “company” (contd.)



(a) Weighted parsing tree of the original sentence





(b) Weighted parsing tree of “small military unit”

In the second gloss “small military unit”, “Large” is the only dependent word of “company” appearing in the dependent word set of “unit”, so the coherence score of gloss 2 is:

$$1.0 \times 1.0 \times 0.8 = 0.8$$

SemEval-2007 Task 07

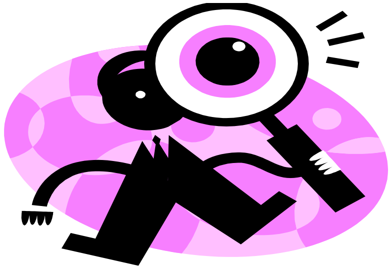
To evaluate the performance of various WSD system, a coarse-grained English all-words task was organized in SemEval 2007 (the Fourth International Workshop on Semantic Evaluations).12 teams

Article	# of words	# of WSD words
a news article about homeless	951	368
a review of the book “Feeding Frenzy”	987	379
an article on traveling in France	1311	500
an article on computer programming	1326	677
a biography of the painter Masaccio	802	345

Experiment result (supervised, unsupervised)

System	Attempted	Precision	Recall	F1
UoR-SSI	100.0	83.21	83.21	83.21
UHD-TreeMatch	100.0	82.68	82.68	82.68
NUS-PT	100.0	82.50	82.50	82.50
NUS-ML	100.0	81.58	81.58	81.58
LCC-WSD	100.0	81.45	81.45	81.45
GPLSI	100.0	79.55	79.55	79.55
UPV-WSD	100.0	78.63	78.63	78.63
TKB-UO	100.0	70.21	70.21	70.21
PU-BCD	90.1	62.80	69.72	66.08
RACAI-SYNWSD	100.0	65.71	65.71	65.71
SUSSZ-FR	72.8	71.73	52.23	60.44
SUSSX-C-WD	72.8	54.54	39.71	45.96
SUSSX-CR	72.8	54.30	39.53	45.75

Motivating Application



[9/11 Flashback: US **Flight Schools** Still Unknowingly Training **...**]

Jul 18, 2012 – More than a decade after the Sept. 11, 2001 **terror** attacks claimed the lives of nearly 3000 Americans, thousands of foreign **flight** students are **...**

[FBI Knew **Terrorists** Were Using **Flight Schools**]

Federal authorities have been aware for years that suspected **terrorists** with ties to Osama bin Laden were receiving **flight training** at schools in the United States **...**

[Homeland Security: Are US **flight schools** still training **terrorists ...**]

Congress is investigating reports that foreign nationals **training** to fly planes in the US were not properly vetted or are in the country on **...**

[Congressional hearing reveals **flight school** security loophole - Los **...**]

U.S. citizens are screened against **terrorism** databases only after **flight training**, when

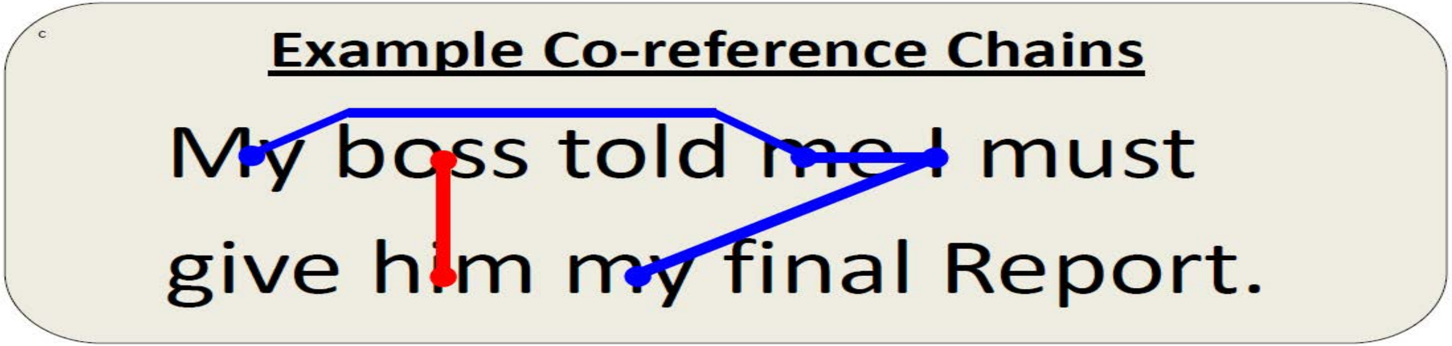
“Muhammad Atta”, “Atta”, “Muhammad”

- Mohamed Atta sent an e-mail to the Academy of Lakeland in Florida, inquiring about flight training.
- On May 17, Mohamed Atta applied for a United States visa.
- Atta arrived on June 3, 2000 at Newark International Airport from Prague.
- Atta began flight training on July 7, 2000 and continued training nearly every day.
- On December 22, Atta and Shehhi applied to Eagle International for large jet.
- On June 27, Atta flew from Fort Lauderdale to Boston, Massachusetts
- **He doesn't really want to learn how to lift off or land.**

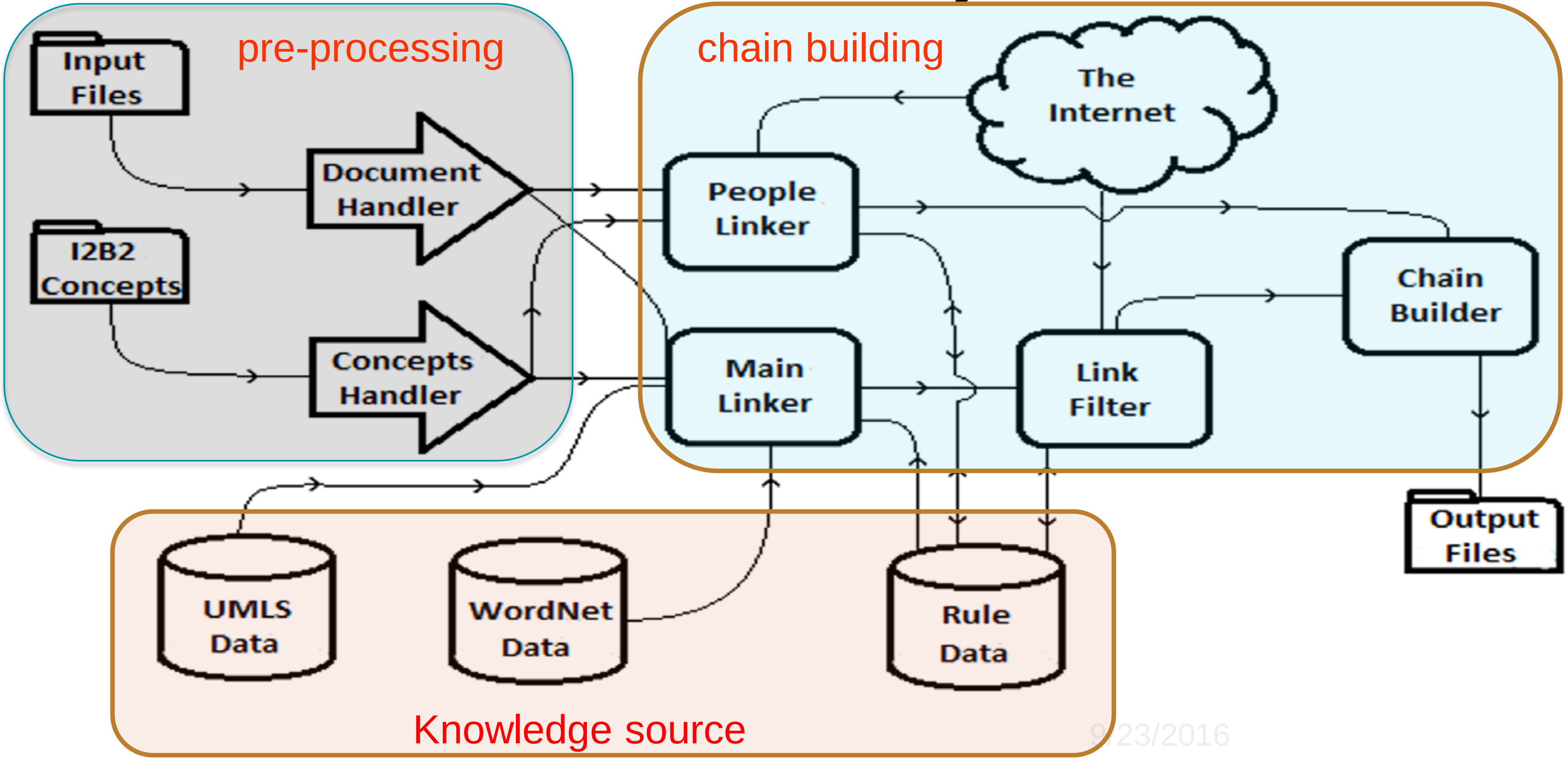
Application: Intelligent Information Retrieval

Co-reference Resolution

Co-reference resolution is the process of linking together concepts that refer to the same entity.

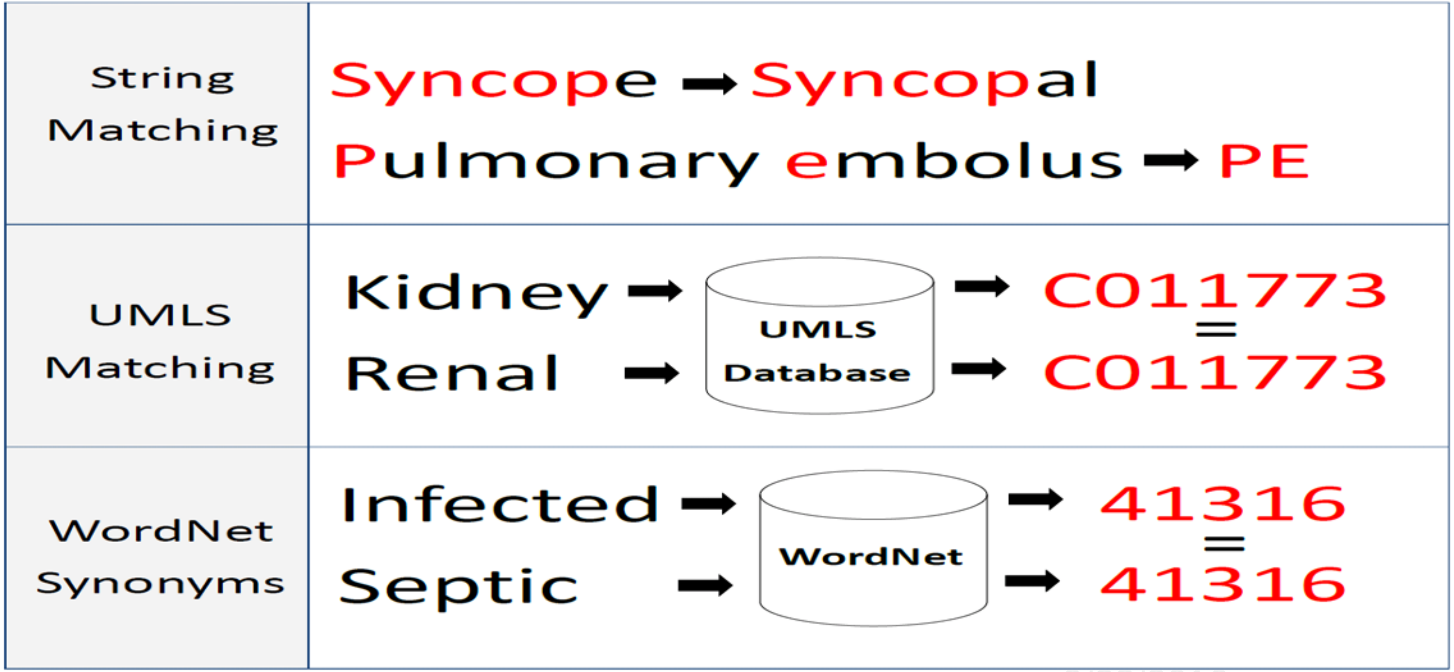


Our co-reference resolution system architecture



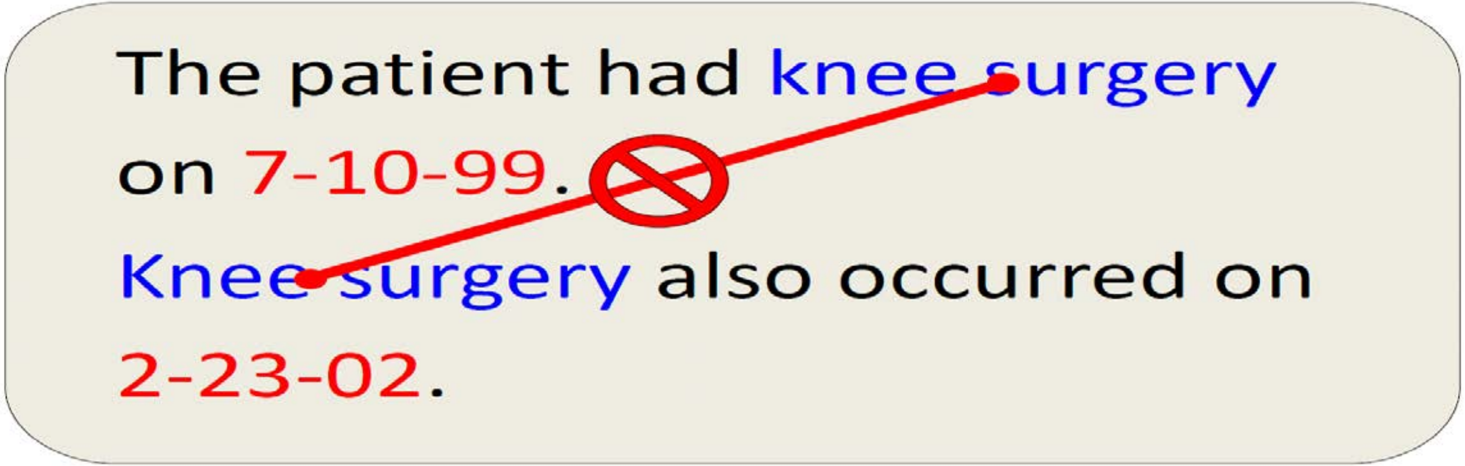
Semantic Rules for Co-reference Resolution

Rules at the lexical semantic level are coded using the UMLS, and WordNet databases to give meaning to the concepts and match the meanings. All pronouns use specific linking rules.



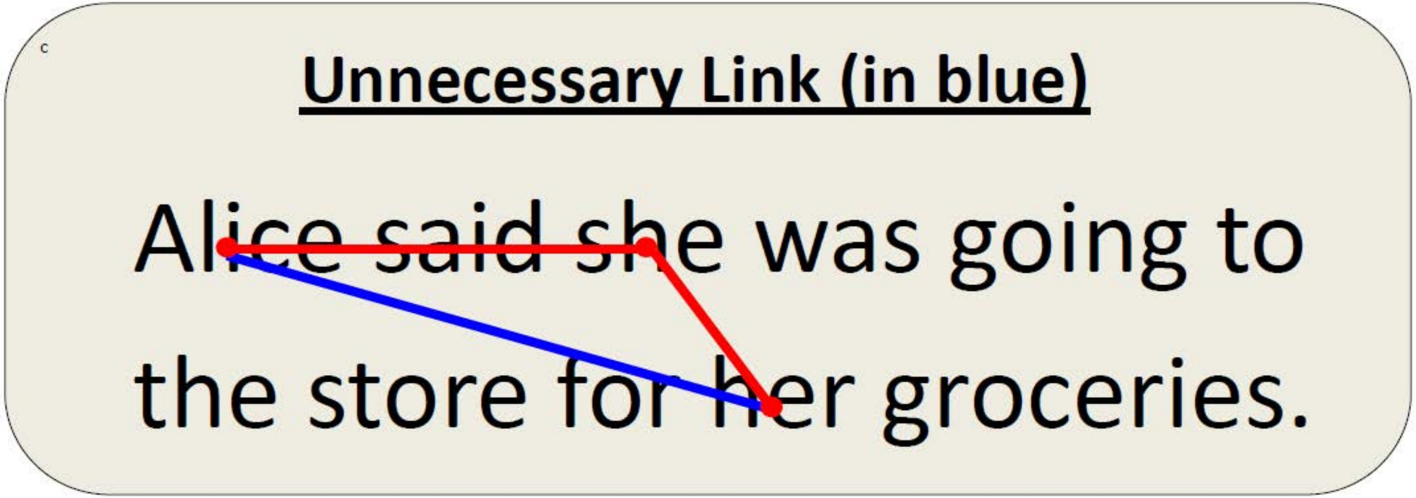
Link Filtering

After linking concepts with the same meaning, links of concepts which do not refer to the same entity must be filtered out. The sentences surrounding the linked concepts are examined for information that indicates if they are different entities. If any relevant information is found and it differs, the link is discarded.



Building Chains

Concepts are first linked in pairs, then, after filtering, unnecessary links are removed to make the chains.



2011 I2B2 Competition

Competition organizers:

- ❑NIH/NLM – 2U54LM008748, Informatics for Integrating Biology and the Bedside (i2b2)
- ❑NIH/NLM – R13 LM010743-01, Shared Task 2010 Analysis of Suicide Notes for Subjective Information
- ❑The VA Consortium for Healthcare Informatics Research (VA HSR HIR 08-374)
- ❑MedQuist Holdings, a provider of integrated clinical documentation solutions
- ❑American Medical Informatics Association

Data

	TRAIN FILES	TEST FILES
Mayo Clinic	**58**	**39**
Clinical reports	30	19
Pathology reports	28	20
Univ. of Pittsburgh Medical Center	**40**	**27**
Discharge reports	10	6
Other reports	9	6
Radiology reports	11	7
Surgical pathology reports	10	8
ODIE	**98**	**66**

Coreference Evaluation

Rank	Team	Unweighted average over MUC, CEAF, and BCUBED P	R	F
1	Microsoft Research Asia	0.906	0.925	0.915
2	Univ. Texas Dallas	0.895	0.918	0.906
3	OPEN Univ.	0.892	0.911	0.901
4	Univ. Houston Downtown	0.895	0.898	0.896
5	HITS gGmbH	0.882	0.894	0.888
6	Brandeis Univ.	0.857	0.915	0.883
7	Centre for Health Informatics, City Univ.	0.895	0.858	0.875
8	Univ. of Illinois at Urbana-Champagne	0.901	0.830	0.861
9	LIMSI-CNRS	0.850	0.862	0.856
10	West Virginia Univ.	0.850	0.846	0.848