

WEB MINING, LANGUAGE PROCESSING AND MACHINE LEARNING: AN INTRODUCTION

R. Basili

(Università di Roma, Tor Vergata)

April 2014

Overview

- Machine Learning, Semantics and NLP
(Trattamento Automatico delle Lingue)
 - Objectives,
 - Methods,
 - Resources and Technologies
 - Applications
- Semantics in Language Processing
- Lexical Semantic tasks and Resources
- Predicate Semantics and Role Labeling
- The role of Tree Kernels
- Conclusions

Speech and Language Processing

- What is S&NLP?
 - To develop programs able to accomplish linguistic tasks, such as:
 - To enable man-machine linguistic interaction
 - Improve communication among people (e.g. MT)
 - Manipulate linguistic objects (ad es. Web pages, documents o telephone calls)
 - Examples:
 - Question Answering
 - Machine Translation
 - Dialogue Agents

Computers, Natural Languages and Applications

- Why *understanding textual contents* by *computers* is useful?
 - Texts are the main carrier of semantic information for many other data types and formats (e.g. multimedia data)
...
 - Natural Language is used to define, transmit, reason and share knowledge (Web is the most evident but not unique example)
 - Information Search is usually based on lexical contents

Processing for *interpretation*

- Processing consists of capture *relevant aspects* of a text
 - Topic (e.g. Politics/Sport)
 - Purposes (e.g. virus/spam in e-mails)
 - People, Organisation or Locations (mentions)
 - Events (e.g. news)
 - Types of communication (e.g. dialogues, planning)
- Result: explicit *representation of the text meaning(s)*
- ... able to trigger some *inferences* (e.g. *relevance*)

Example: News agency



The screenshot shows the ANSA.IT website interface. At the top left is the logo "ansa.it" with the tagline "IL PORTALE DELL'INFORMAZIONE". To the right are navigation links for "Sitemap", "Prodotti", and "Contattaci". Below this is a horizontal menu with categories: "Home", "Italia", "Mondo", "Società", "Internet", "Economia", "Sport", and "Spettacolo". A sub-menu titled "I FATTI DEL GIORNO" is visible, with a selected item: "RUGBY: GRAND'ITALIA AL FLAMINIO, SCOZIA BATTUTA 20-14".

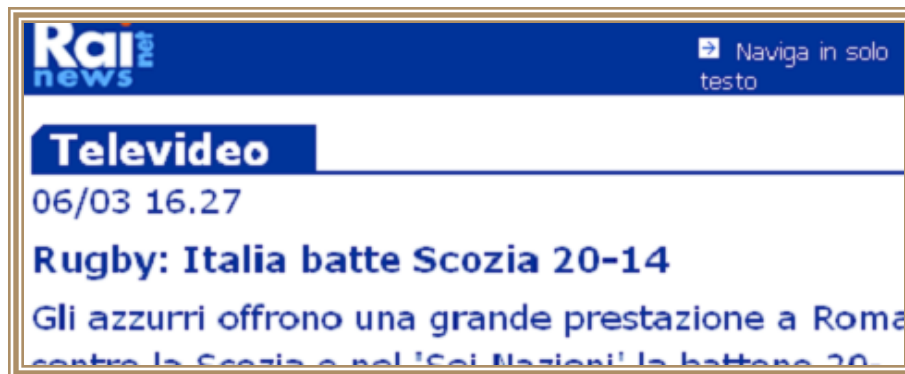
The main article features a photograph of a rugby player in action, wearing a white jersey, reaching for a ball. The headline reads: "RUGBY: GRAND'ITALIA AL FLAMINIO, SCOZIA BATTUTA 20-14". The text of the article begins: "ROMA - L'Italia ha battuto la Scozia per 20-14 (9-9) in una partita della terza giornata del torneo delle Sei Nazioni 2004. Evitato così 'il cucchiaino di legno' che spetta alla squadra".

Below the headline, there are two paragraphs of text. The first paragraph discusses the importance of the match for the Italian team, mentioning John Kirwan and the goalkeeping situation. The second paragraph provides context on the team's performance against Scotland, noting a previous victory in 1998.

On the left side of the page, there is a vertical sidebar with various links and categories, including "eti", "ca News Multi", "ANALI E RUBR", "n edicola ogg", "n libro al gio", "leteo", "orsa & Financ", "rasporti e", "nfrastuttur", "ambiente", "eni Culturali Sicilia", ".grealimentare azio", ".grealimentare icilia", "enova 2004", "orino 2006", "ews locali", "loda", "oto e lotterie", "azzetta Ufficiale", "onale Turismo", and "TRASPORTI INFRASTRUTTURE".

News (2)

- Requirements of a correct interpretation are (at least):
 - “ha battuto” is the main verb
 - ... used as a transitive verb
 - “sport” meaning (no one is beaten/hit here!)
 - Italia and Scozia are the grammatical subject and object respectively (☺)
 - Italia is not the country , but a team (!), (as well as Scozia)
 - giornata is the turn and NOT the day
- Many other equivalent linguistic forms e.g.



News (3): Multilinguality

BBC CATEGORIES TV RADIO COMMUNICATE WHERE I LIVE INDEX SEARCH Go

Low Graphics version | Change Edition Make this my homepage | Help

BBC SPORT **RUGBY UNION** WATCH/LISTEN TO BBC SPORT OPEN AUDIO/VIDEO CONSOLE

Sport Homepage

Rugby Union

Six Nations

Live Scores

Results

Fixtures

Kicking Kings

Wallpaper

TV trails

Your Say:

Scrum V >>

Academy >>

Rules >>

Skills >>

Daily E-mail

Mobiles

Fun and Games

Question of Sport

CHOOSE A SPORT

Select

Go

BBC NEWS

BBC WEATHER

ACADEMY

BBC SPORT

Last Updated: Saturday, 6 March, 2004, 15:18 GMT

[E-mail this to a friend](#)

[Printable version](#)

Italy 20-14 Scotland

Scotland look set for the Six Nations wooden spoon after being convincingly outplayed by Italy.

Hooker Fabio Ongaro scored Italy's only try - at the start of the second half - although replays showed he had failed to ground it.

Despite that, the scoreline flattered the Scots, who struggled to match the Italians in the set pieces.

Simon Webster scored a late consolation try and the remaining points came from Chris Paterson and Roland de Marigny.

The game lived up to its reputation as the battle between the Six Nations' two worst sides from the opening whistle.

Neither team showed any invention on the attack, instead generally opting to kick the ball or keep it in the forwards.

But Italy were the far more assured on the basics and stronger in the set pieces, with De Marigny kicking well from the spot as well as out of hand in his first start at fly-half.



• **Italy 20 (9)**
Try: Ongaro
Pens: De Marigny (5)
• **Scotland 14 (9)**
Try: Webster
Pens: Paterson (3)

[All the action as it happened](#)
[Match photos](#)

WATCH AND LISTEN

Scotland coach Matt Williams

"Our mistake rate was unacceptable"

[VIDEO](#)

RBS SIX NATIONS 2004

[Latest international news](#)

WEEKEND THREE

[Ireland stun England](#)

[Italy 20-14 Scotland](#)

[Wales v France](#)

IN VIDEO

[OPEN](#) Six Nations TV highlights

[VIDEO](#) Six Nations Forum

PLAYER PUNDITS

[Matt Dawson](#)

[Iestyn Harris](#)

[Brian O'Driscoll](#)

[Gordon Bulloch](#)

BBC COVERAGE

[Six Nations on the BBC](#)

OFFICIAL WEBSITE

[RBS Six Nations 2004](#)

The BBC is not responsible for the content of external internet sites

RELATED INTERNET LINKS:

[Scottish Rugby](#)

The BBC is not responsible for the content of external internet sites

ALSO IN THIS SECTION

[Ireland stun England](#)

[Italy upset Scotland](#)

[Traillie fit to face Wales](#)

[Wales v France](#)

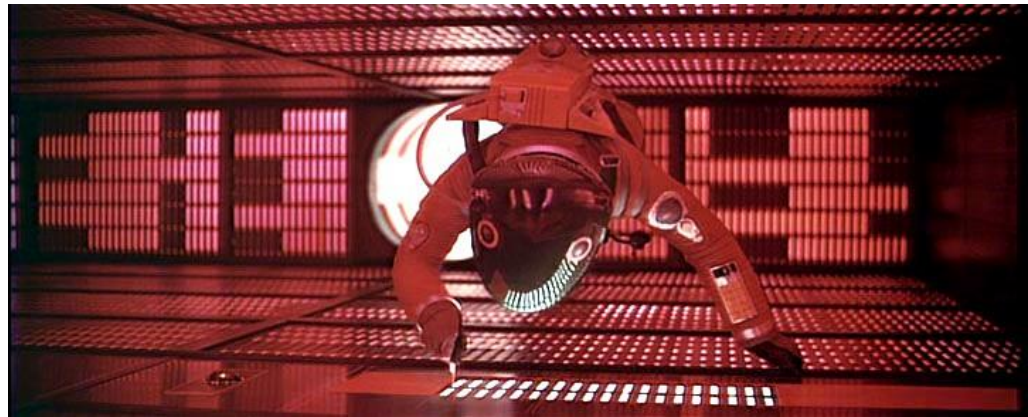
Which knowledge?

- HAL 9000, da “*2001: A Space Odyssey*”
- Dave: *Open the pod bay doors, Hal.*
- HAL: *I’m sorry Dave, I’m afraid I can’t do that.*



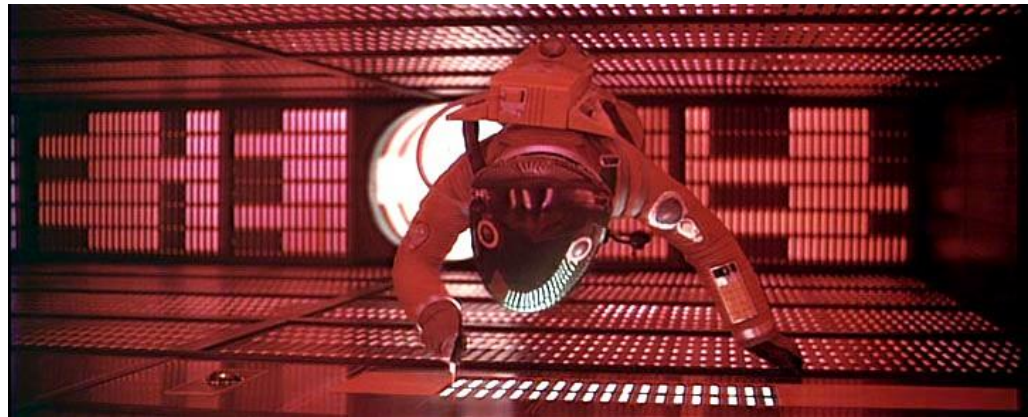
What's HAL knowledge?

- Speech Recognition and Synthesis
 - (Phoneme) Dictionary
 - Phonetics (how to recognize/produce English sounds)
- Language Understanding
 - English lexical knowledge,
 - What words mean
 - How they can be combined (What is a `pod bay door'?)
 - Knowledge about the syntagmatic structure
 - I'm I do, Sorry that afraid Dave I'm can't



What's HAL knowledge? (2)

- Dialogue and pragmatics
 - “open the door” is a request (not a statement or an information search)
 - What does `that' mean in `I can't do that'?
 - Answering is a gentle reaction even if you're planning to kill.
 - It is better to show a cooperative attitude (I'm afraid, I can't...)
- Even an automatic airflight booking system asks more or less the same knowledge



Question Answering

- Cosa significa “porta”?
- In quale anno e' nato Mozart?
- Quante erano le provincie italiane sino al 1995?
- C'era uno sconto sull'acquisto dei libri di IA da Amazon ieri?
- Cosa pensano gli scienziati riguardo alla legalizzazione della clonazione?

Some reflections

- Understanding linguistic objects requires knowledge about:
 - The language (e.g. *syntax*)
 - the world (e.g. *rugby, teams and countries*)
 - How the first make reference to the second
- Intelligent Access and Publication requires knowledge about:
 - The purpose, i.e. search vs. command
 - The world in which the communication is immerse
 - *Text producers vs. text users*

Previous experiences: QA @ RTV, the Know All system

The screenshot shows a web browser window displaying the Know-All application. The browser's address bar shows the URL `http://10.6.0.103:3320/Know_All/index.html`. The page has a navigation menu with links for "Architecture", "How To", "Know-All Service", "Semantic Role Labeling Service", and "Credits". The main heading "Know-All" is displayed in a large, multi-colored font, with the version number "1.0b" to its right. Below the heading is a text input area containing the questions "Where is Taj Mahal?" and "What is a tsunami?". Underneath the input area are three sets of radio buttons for configuration: "Semantic Role Labeling" (disabled: selected, enabled: unselected), "Search Engine I" (selected) and "Search Engine II" (unselected), and "Identifinder" (selected) and "Chaos" (unselected). There are "Submit" and "Clear" buttons below the configuration options. At the bottom, a scrollable output area shows the results: "Answer 0001 (offers by Identifinder)" followed by a list of three items: "0001: India", "0002: Agra", and "0003: Brooklyn".

Architecture How To **Know-All Service** Semantic Role Labeling Service Credits

Know-All

1.0b

Where is Taj Mahal?
What is a tsunami?

Semantic Role Labeling: disabled: enabled:
Search Engine I: Search Engine II:
Identifinder: Chaos:

Answer 0001 (offers by Identifinder)

0001: India
0002: Agra
0003: Brooklyn

Know-All

1.0b

Where is Taj Mahal?
What is a tsunami?

Semantic Role Labeling: disabled: enabled:

Search Engine I: Search Engine II:

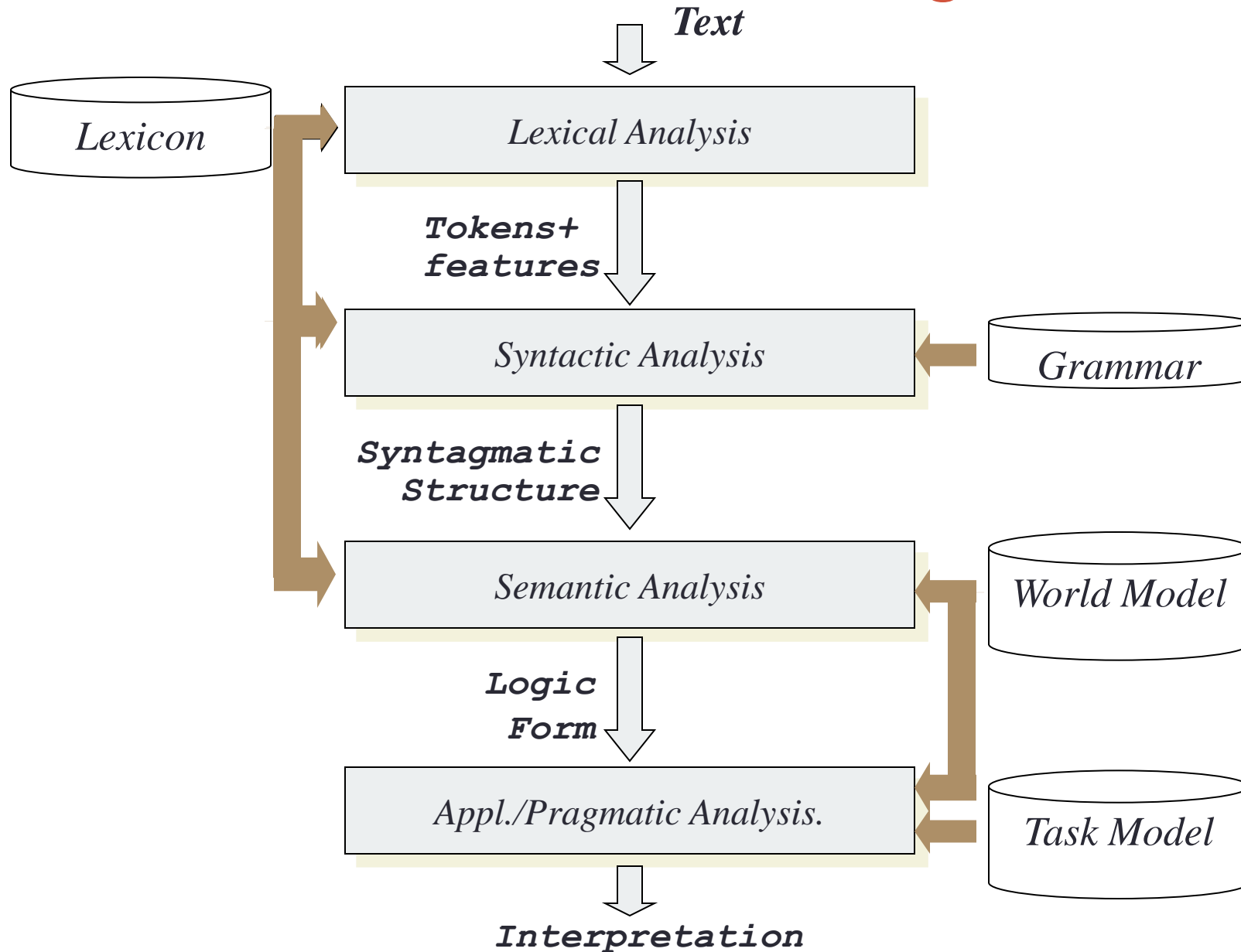
Identifinder: Chaos:

Answer 0002 (offers by AM)

0001: In Japan, tsunamis are considered almost as big a threat as earthquakes on land.
0002: A tsunami is a massive wave caused by an earthquake or volcanic eruption.
0003: David called it a tsunami of new opportunity, he said.

Know-All status: idle

Nature and models of linguistic data



Language Study: a computational perspective

- Main questions for the research in linguistics:
- What does it mean to know the mother tongue? (Competence)
- How language is used? (Performance)
- How knowledge of language is acquired? (Language Acquisition)
- How knowledge of language is represented in the brain?

Syntax and Semantics in textual data

- **Compositionality**

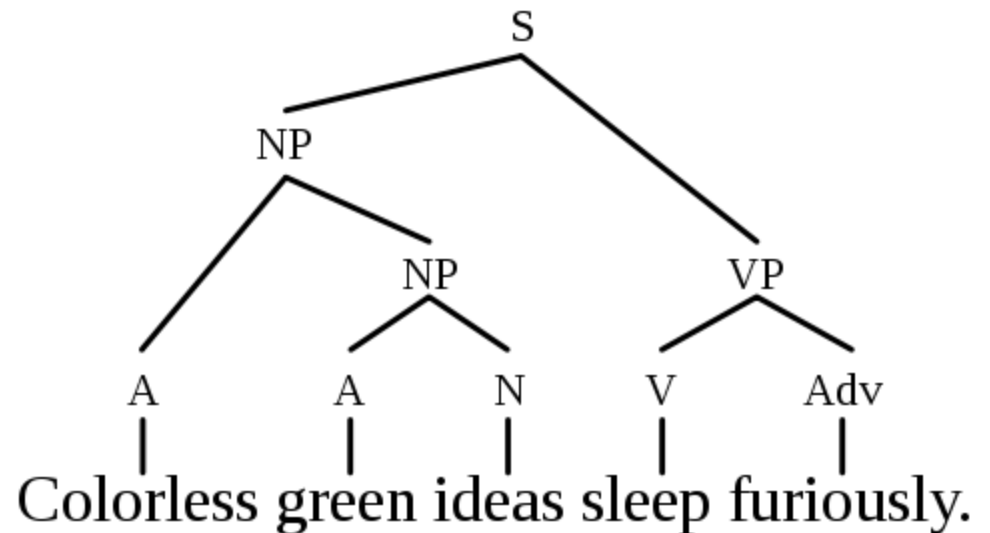
- The meaning of a complex expression is solely determined by the meanings of its constituent expressions and the rules used to combine them.
- *"I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements. All natural languages are languages in this sense. Similarly, the set of "sentences" of some formalized system of mathematics can be considered a language"* Chomsky 1957

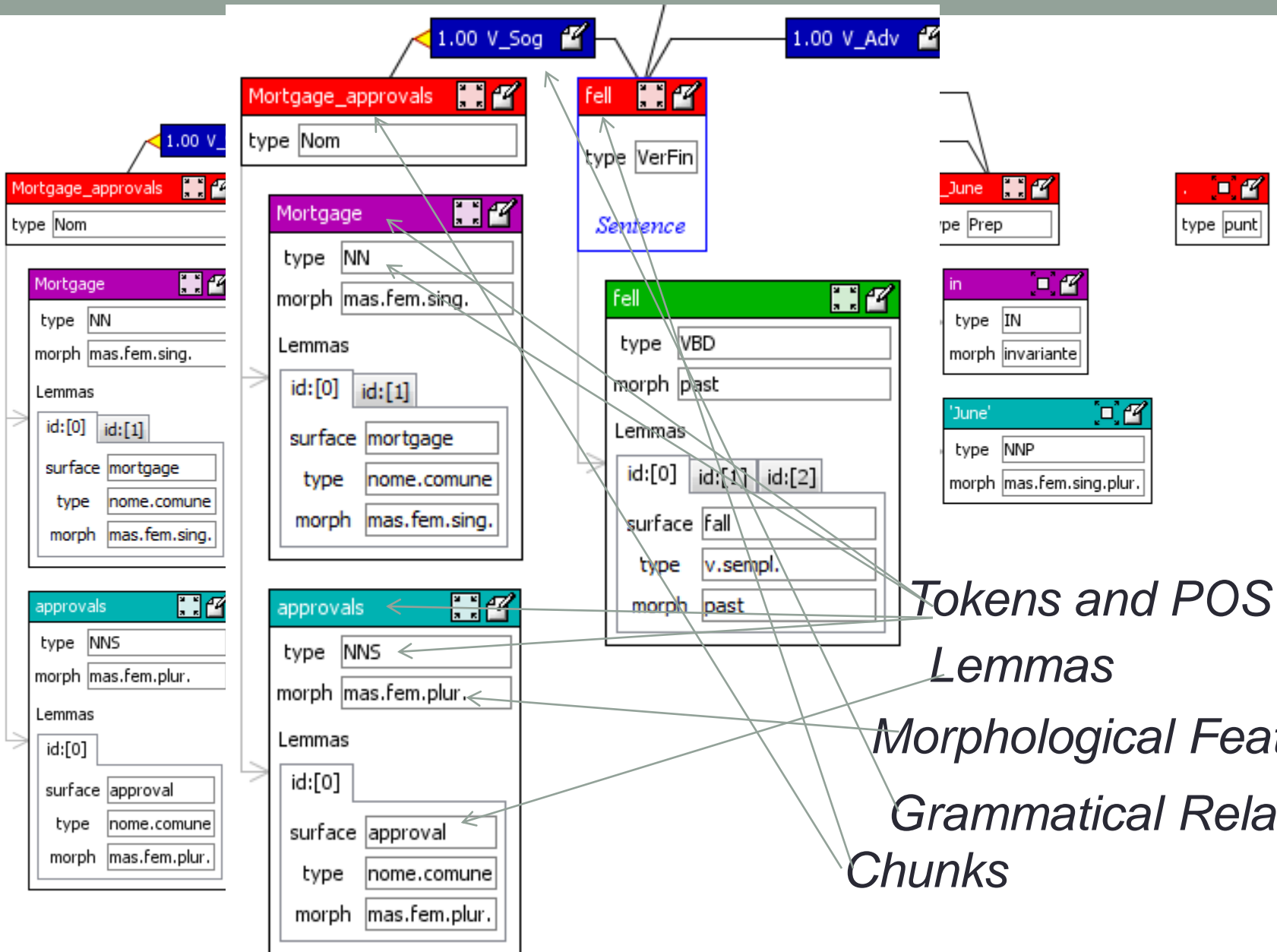
Syntax

- In linguistics, **syntax** is the study of the rules that govern the structure of sentences, and which determine their relative grammaticality.
- Such rules govern a number of language phenomena as systems for phonology, morphology, syntax as well as discourse

Syntax, Grammars and Trees

1. S \rightarrow NP VP
2. S \rightarrow NP
3. NP \rightarrow PN
4. NP \rightarrow N
5. NP \rightarrow Adj N





Tokens and POS tags
Lemmas
Morphological Features
Grammatical Relations
Chunks

FT (July, 29): *Mortgage approvals fell sharply in June.*

Overview

- Machine Learning, Semantics and NLP
(Trattamento Automatico delle Lingue)
 - Objectives,
 - Methods,
 - Resources and Technologies
 - Applications
- Semantics in Language Processing
- Lexical Semantic tasks and Resources
- Predicate Semantics and Role Labeling
- The role of Tree Kernels
- Conclusions

NLP: the semantic level

Ambiguity

- *Gianni observed the girl with the binocular (ambiguous)*
- *Gianni observed her with the binocular (non ambiguous)*
- *Gianni already knew the girl with the binocular (non amb.)*

- *Every man loves his mother (ambiguous)*
- *His mother loves every man (non ambigua)*

Synonymy and variability

- *Gianni helped Piero*
- *Piero has been helped by Gianni (synonymous)*
- *Piero helped Gianni
(non synonymous)*
- *Red party vs. Red apple*
- *Red Party vs. Communist Party*

Inconsistency

- *# Gianni killed the dog, that never died ...*
- *# Yesterday morning I will wake up at 7pm (...)*
- *Colorless green ideas sleep furiously*

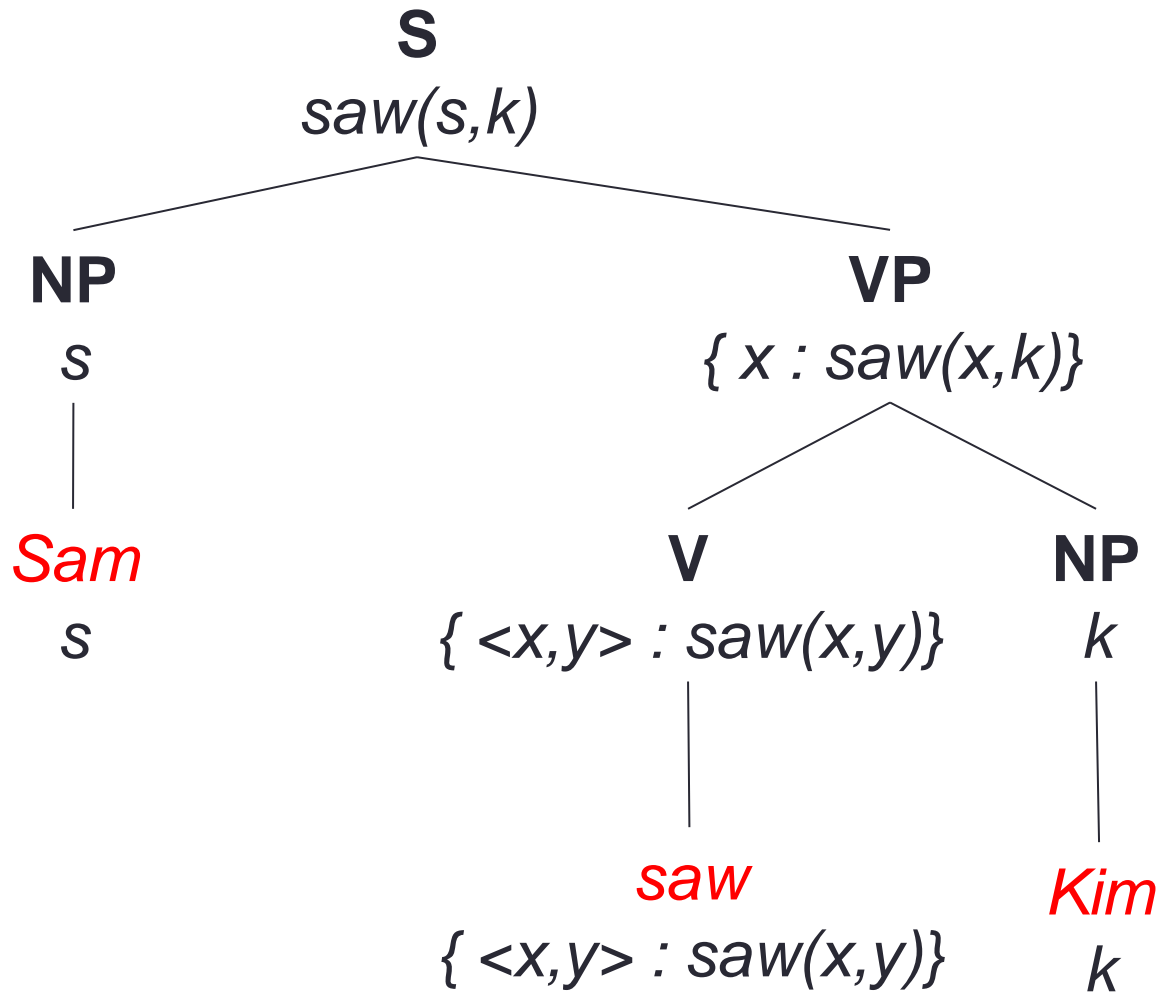
Semantics

- For the sentence:

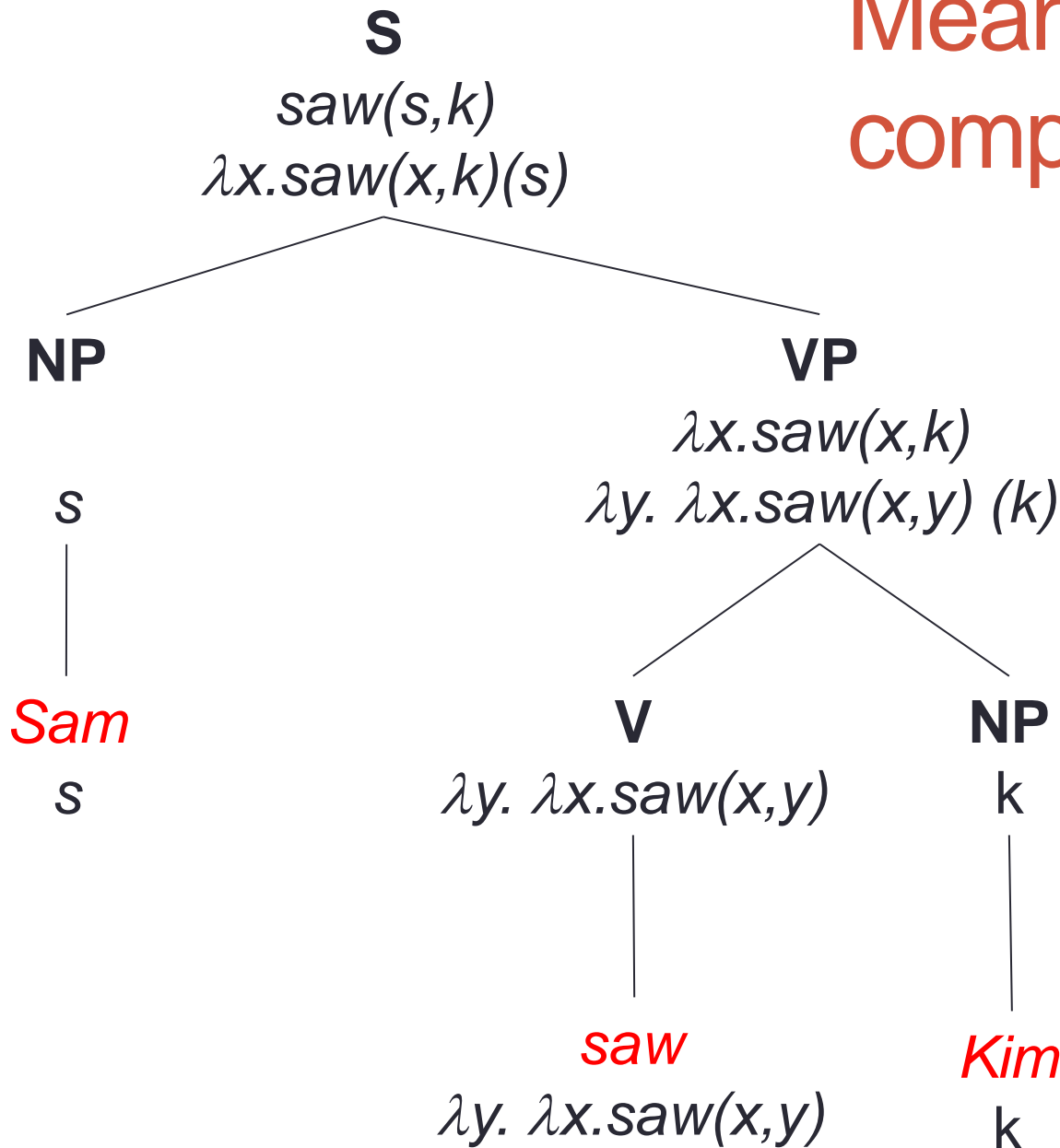
John saw Kim

- What about its **meaning**?
- Properties:
 - **It must be derivable compositionally**, i.e. from the meanings of the individual constituents, i.e. *Kim*, *John* and *see*
 - **Independence on syntactic phenomenon**, e.g. *Kim was seen by John*
 - It must **support inferences**
 - Who was seen by John?
 - John saw Kim. He started running to her.

Truth conditional view on meaning



Meaning as a computation



Semantics

- Words *senses* activates predicates
 - *Bank/money* vs. *bank/river*
 - Usually in the lexicon:
 - **bank_1 (X)** VS. **Bank_2 (X)**
- Verbs are predicates that express:
 - Events/states as complex relationships among participants
 - John gave Mary a book
 - John gave a book to Mary
 - John was running on the hill

Three Perspectives on Meaning

1. **Lexical Semantics**

- The meanings of **individual words**

2. **Formal Semantics** (or **Compositional Semantics** or **Sentential Semantics**)

- How those meanings combine to make meanings for **individual sentences or utterances**

3. **Discourse or Pragmatics**

- How those meanings combine with each other and with other facts about various kinds of context to make meanings for a **text or discourse**
- **Dialog or Conversation** is often lumped together with Discourse

Overview

- Machine Learning, Semantics and NLP
(Trattamento Automatico delle Lingue)
 - Objectives,
 - Methods,
 - Resources and Technologies
 - Applications
- Semantics in Language Processing
- Lexical Semantic tasks and Resources
- Predicate Semantics and Role Labeling
- The role of Tree Kernels
- Conclusions

Relationships between word meanings

- Homonymy
- Polysemy
- Synonymy
- Antonymy
- Hypernymy
- Hyponymy
- Meronymy

Homonymy

- **Homonymy:**
 - Lexemes that share a form
 - Phonological, orthographic or both
 - But have unrelated, distinct meanings
 - Clear example:
 - Bat (wooden stick-like thing) vs
 - Bat (flying scary mammal thing)
 - Or bank (financial institution) versus bank (riverside)
 - Can be also homophones, homographs, or both:
 - Homophones:
 - *Write* and *right*
 - *Piece* and *peace*

Polysemy

- *The **bank** is constructed from red brick
I withdrew the money from the **bank***
- Are those the same sense?
- Or consider the following WSJ example
 - *While some banks furnish sperm only to married women, others are less restrictive*
- Which sense of *bank* is this?
 - Is it distinct from (homonymous with) the river bank sense?
 - How about the *savings bank* sense?

Polysemy

- A single lexeme with multiple **related** meanings (*bank* the building, *bank* the financial institution) is polysemous
- Most non-rare words have multiple meanings
 - The number of meanings is related to frequency in the texts
 - Verbs tend more to polysemy
 - Distinguishing polysemy from homonymy isn't always easy (and even necessary)

... in Wordnet (Miller, 1991)

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	<i>breakfast</i> ¹ → <i>meal</i> ¹
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> ¹ → <i>lunch</i> ¹
Member Meronym	Has-Member	From groups to their members	<i>faculty</i> ² → <i>professor</i> ¹
Has-Instance		From concepts to instances of the concept	<i>composer</i> ¹ → <i>Bach</i> ¹
Instance		From instances to their concepts	<i>Austen</i> ¹ → <i>author</i> ¹
Member Holonym	Member-Of	From members to their groups	<i>copilot</i> ¹ → <i>crew</i> ¹
Part Meronym	Has-Part	From wholes to parts	<i>table</i> ² → <i>leg</i> ³
Part Holonym	Part-Of	From parts to wholes	<i>course</i> ⁷ → <i>meal</i> ¹
Antonym		Opposites	<i>leader</i> ¹ → <i>follower</i> ¹

WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	<i>fly</i> ⁹ → <i>travel</i> ⁹
Troponym	From a verb (event) to a specific manner elaboration of that verb	<i>walk</i> ¹ → <i>stroll</i> ¹
Entails	From verbs (events) to the verbs (events) they entail	<i>snore</i> ¹ → <i>sleep</i> ¹
Antonym	Opposites	<i>increase</i> ¹ ⇔ <i>decrease</i> ¹

WordNet Hierarchies

Sense 3

bass, basso --

(an adult male singer with the lowest voice)

- => singer, vocalist, vocalizer, vocaliser
- => musician, instrumentalist, player
- => performer, performing artist
- => entertainer
- => person, individual, someone...
- => organism, being
- => living thing, animate thing,
- => whole, unit
- => object, physical object
- => physical entity
- => entity
- => causal agent, cause, causal agency
- => physical entity
- => entity

Sense 7

bass --

(the member with the lowest range of a family of musical instruments)

- => musical instrument, instrument
- => device
- => instrumentality, instrumentation
- => artifact, artefact
- => whole, unit
- => object, physical object
- => physical entity
- => entity

How is “sense” defined in WordNet?

- The set of near-synonyms for a WordNet sense is called a **synset (synonym set)**; it’s their version of a sense or a concept
- Example: **chump** as a noun to mean
 - ‘a person who is gullible and easy to take advantage of’
`{chump1, fool2, gull1, mark9, patsy1, fall guy1, sucker1, soft touch1, mug2}`
- Each of these senses share this same gloss
- Thus for WordNet, the meaning of this sense of **chump** is this list.

I sensi in Wordnet

- L'unita' fondamentale della organizzazione lessicale di Wordnet e' il synonymy set, o synset
- Un synset e' formato da un insieme di parole che (secondo un certo aspetto del loro significato) sono sinonimi

1. set, circle, band, lot -- (an unofficial association of people or groups; "the smart set goes there"; ...)
2. band -- (instrumentalists not including string players)
3. band, stria, striation -- (a stripe of contrasting color; "chromosomes exhibit characteristic bands")
4. band, banding, stripe -- (a strip or stripe of a contrasting color or material)
5. dance band, band, dance orchestra -- (a group of musicians playing popular music for dancing)
6. band -- (a range of frequencies between two limits)
7. band -- (something elongated that is worn around the body or one of the limbs)
8. ring, band -- (jewelry consisting of a circular band of a precious metal worn on the finger; "she had rings on every finger")
9. band -- (put around something to hold it together)

Wordnet Size

Wordnet (1.7): Scala:

POS	Unique Strings	Synsets	(W,Sense) Pairs
Noun	109,195	75,804	134,716
Verb	11,088	13,214	24,169
Adjective	21,460	18,576	31,184
Adverb	4,607	3,629	5,748
Totals	146,350	111,223	195,817

Wordnet polisemy

Wordnet (1.7): Polysemy information:

POS	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	94,685	14,510	40,002
Verb	5,920	5,168	18,221
Adjective	15,981	5,479	15,175
Adverb	3,820	787	1,900
Totals	120,406	25,944	75,298

Wordnet Avg polisemy

Wordnet: Average polisemy (AvPol) :

POS	Including Monosemous Words	Excluding Monosemous Words
Noun	1.23	2.75
Verb	2.17	3.52
Adjective	1.45	2.76
Adverb	1.24	2

Word Similarity

- Synonymy is a binary relation
 - Two words are either synonymous or not
- We want a looser metric
 - Word similarity or
 - Word distance
- Two words are more similar
 - If they share more features of meaning

Word Similarity

- Actually these are really relations between **senses**:
 - Instead of saying “*bank is like fund*”
 - We say
 - Bank1 *is similar to* fund3
 - Bank2 *is similar to* slope5
- Similarity are computed over both words and senses

Why word similarity

- Spell Checking
- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading

WSD: Practical Applications

- Machine Translation
 - Translate “bill” from English to Spanish
 - Is it a “pico” or a “cuenta”?
 - Is it a bird jaw or an invoice?
- Information Retrieval
 - Find all Web Pages about “cricket”
 - The sport or the insect?
- Question Answering
 - What is George Miller’s position on gun control?
 - The psychologist or US congressman?
- Knowledge Acquisition
 - Add to KB: Herb Bergson is the mayor of Duluth.
 - Minnesota or Georgia?

Word Sense Disambiguation: Overview of the Problem

- Many words have several meanings (homonymy / polysemy)

–Ex: “chair” – furniture or person

–Ex: “child” – young person or human offspring

- Determine which sense of a word is used in a specific sentence

- **Note:**

- often, the different senses of a word are closely related

- Ex: **title** – right of legal ownership

- – document that is evidence of the legal ownership,

- sometimes, several senses can be “activated” in a single context (co-activation)

- Ex: “*This could bring competition to the trade*”

- **competition:** – the act of competing

- – the people who are competing

Word Senses

- The *meaning* of a word in a given context
- Word sense representations
 - With respect to a dictionary

chair = a seat for one person, with a support for the back; "he put his coat over the back of the chair and sat down"

chair = the position of professor; "he was awarded an endowed chair in economics"

- With respect to the translation in a second language

chair = chaise

chair = directeur

- With respect to the context where it occurs (discrimination)

"Sit on a *chair*" "Take a seat on this *chair*"

"The *chair* of the Math Department" "The *chair* of the meeting"

Approaches to Word Sense Disambiguation

- Knowledge-Based Disambiguation
 - use of external lexical resources such as dictionaries and thesauri
 - discourse properties
- Supervised Disambiguation
 - based on a labeled training set
 - the learning system has:
 - a training set of feature-encoded inputs AND
 - their appropriate sense label (category)
- Unsupervised Disambiguation
 - based on unlabeled corpora
 - The learning system has:
 - a training set of feature-encoded inputs BUT
 - NOT their appropriate sense label (category)

All Words Word Sense Disambiguation

- Attempt to disambiguate all open-class words in a text

“He **put** his **suit** over the **back** of the **chair**”

- Knowledge-based approaches
- Use information from dictionaries
 - Definitions / Examples for each meaning
 - Find similarity between definitions and current context
- Position in a semantic network
 - Find that “**table**” is closer to “**chair/furniture**” than to “**chair/person**”
- Use discourse properties
 - A word exhibits the same sense in a discourse / in a collocation

All Words Word Sense Disambiguation

- Minimally supervised approaches
 - Learn to disambiguate words using small annotated corpora
 - E.g. SemCor – corpus where all open class words are disambiguated
 - 200,000 running words
- Most frequent sense

Targeted Word Sense Disambiguation

- Disambiguate one target word

“Take a seat on this chair”

“The chair of the Math Department”

- WSD is viewed as a typical classification problem
 - use machine learning techniques to train a system
- Training:
 - Corpus of occurrences of the target word, each occurrence annotated with appropriate sense
 - Build feature vectors:
 - a vector of relevant linguistic features that represents the context (ex: a window of words around the target word)
- Disambiguation:
 - Disambiguate the target word in new unseen text

Targeted Word Sense Disambiguation

- Take a window of n word around the target word
- Encode information about the words around the target word
 - typical features include: words, root forms, POS tags, frequency, ...
 - An electric guitar and bass player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.
 - Surrounding context (local features)
 - [(guitar, NN1), (and, CJC), (player, NN1), (stand, VVB)]
 - Frequent co-occurring words (topical features)
 - [*fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band*]
 - [0,0,0,1,0,0,0,0,0,0,1,0]
 - Other features:
 - [followed by "player", contains "show" in the sentence,...]
 - [yes, no, ...]

Unsupervised Disambiguation

- Disambiguate word senses:
 - without supporting tools such as dictionaries and thesauri
 - without a labeled training text
- Without such resources, word senses are not *labeled*
 - We cannot say “**chair/furniture**” or “**chair/person**”
- We can:
 - Cluster/group the contexts of an ambiguous word into a number of groups
 - *Discriminate* between these groups without actually labeling them

Unsupervised Disambiguation

- Hypothesis: same senses of words will have similar neighboring words
- Disambiguation algorithm
 - Identify context vectors corresponding to all occurrences of a particular word
 - Partition them into regions of high density
 - Assign a sense to each such region

“Sit on a **chair**”

“Take a seat on this **chair**”

“The **chair** of the Math Department”

“The **chair** of the meeting”

Evaluating Word Sense Disambiguation

- Metrics:
 - Precision = percentage of words that are tagged correctly, out of the words addressed by the system
 - Recall = percentage of words that are tagged correctly, out of all words in the test set

- Example

- Test set of 100 words
- System attempts 75 words
- Words correctly disambiguated 50

$$\text{Precision} = 50 / 75 = 0.66$$

$$\text{Recall} = 50 / 100 = 0.50$$

- Special tags are possible:
 - Unknown
 - Proper noun
 - Multiple senses
- Compare to a gold standard
 - SEMCOR corpus, SENSEVAL corpus, ...

Evaluating Word Sense Disambiguation

- Difficulty in evaluation:
 - Nature of the senses to distinguish has a huge impact on results
- Coarse versus fine-grained sense distinction

chair = a **seat** for one person, with a support for the back; "he put his coat over the back of the chair and sat down"

chair = the position of **professor**; "he was awarded an endowed chair in economics"

bank = a **financial institution** that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home"

bank = a **building** in which commercial banking is transacted; "the bank is on the corner of Nassau and Witherspoon"

- Sense maps
 - Cluster similar senses
 - Allow for both fine-grained and coarse-grained evaluation

Bounds on Performance

- Upper and Lower Bounds on Performance:
 - Measure of how well an algorithm performs relative to the difficulty of the task.
- Upper Bound:
 - Human performance
 - Around 97%-99% with few and clearly distinct senses
 - Inter-judge agreement:
 - With words with clear & distinct senses – 95% and up
 - With polysemous words with related senses – 65% – 70%
- Lower Bound (or baseline):
 - The assignment of a random sense / the most frequent sense
 - 90% is excellent for a word with 2 equiprobable senses
 - 90% is trivial for a word with 2 senses with probability ratios of 9 to 1

References

- (Gale, Church and Yarowsky 1992) Gale, W., Church, K., and Yarowsky, D. *Estimating upper and lower bounds on the performance of word-sense disambiguation programs* ACL 1992.
- (Miller et. al., 1994) Miller, G., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. *Using a semantic concordance for sense identification*. ARPA Workshop 1994.
- (Miller, 1995) Miller, G. Wordnet: A lexical database. ACM, 38(11) 1995.
- (Senseval) Senseval evaluation exercises <http://www.senseval.org>

Lesk Algorithm

- (Michael Lesk 1986): Identify senses of words in context using definition overlap

Algorithm:

1. Retrieve from MRD all sense definitions of the words to be disambiguated
2. Determine the definition overlap for all possible sense combinations
3. Choose senses that lead to highest overlap

Example: disambiguate PINE CONE

- PINE
 1. kinds of evergreen tree with needle-shaped leaves
 2. waste away through sorrow or illness
- CONE
 1. solid body which narrows to a point
 2. something of this shape whether solid or hollow
 3. fruit of certain evergreen trees

$$\text{Pine\#1} \cap \text{Cone\#1} = 0$$

$$\text{Pine\#2} \cap \text{Cone\#1} = 0$$

$$\text{Pine\#1} \cap \text{Cone\#2} = 1$$

$$\text{Pine\#2} \cap \text{Cone\#2} = 0$$

$$\text{Pine\#1} \cap \text{Cone\#3} = 2$$

$$\text{Pine\#2} \cap \text{Cone\#3} = 0$$

Lesk Algorithm for More than Two Words?

- *I saw a man who is 98 years old and can still walk and tell jokes*
 - nine open class words: *see(26), man(11), year(4), old(8), can(5), still(4), walk(10), tell(8), joke(3)*
- 43,929,600 sense combinations! How to find the optimal sense combination?
- Simulated annealing (Cowie, Guthrie, Guthrie 1992)
 - Define a function E = combination of word senses in a given text.
 - Find the combination of senses that leads to highest definition overlap (*redundancy*)
 1. Start with E = the most frequent sense for each word
 2. At each iteration, replace the sense of a random word in the set with a different sense, and measure E
 3. Stop iterating when there is no change in the configuration of senses

Lesk Algorithm: A Simplified Version

- Original Lesk definition: measure overlap between sense definitions for all words in context
 - Identify simultaneously the correct senses for all words in context
- Simplified Lesk (Kilgarriff & Rosensweig 2000): measure overlap between sense definitions of a word and current context
 - Identify the correct sense for one word at a time
- Search space significantly reduced

Lesk Algorithm: A Simplified Version

- Algorithm for simplified Lesk:
 1. Retrieve from MRD all sense definitions of the word to be disambiguated
 2. Determine the overlap between each sense definition and the current context
 3. Choose the sense that leads to highest overlap

Example: disambiguate PINE in

“Pine cones hanging in a tree”

- PINE

1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness

Pine#1 \cap Sentence = 1
Pine#2 \cap Sentence = 0

Evaluations of Lesk Algorithm

- Initial evaluation by M. Lesk
 - 50-70% on short samples of text manually annotated set, with respect to Oxford Advanced Learner's Dictionary
- Simulated annealing
 - 47% on 50 manually annotated sentences
- Evaluation on Senseval-2 all-words data, with back-off to random sense (Mihalcea & Tarau 2004)
 - Original Lesk: 35%
 - Simplified Lesk: 47%
- Evaluation on Senseval-2 all-words data, with back-off to most frequent sense (Vasilescu, Langlais, Lapalme 2004)
 - Original Lesk: 42%
 - Simplified Lesk: 58%

Yarowsky Algorithm

- (Yarowsky 1995)
- Similar to co-training
- Differs in the basic assumption (Abney 2002)
 - “view independence” (co-training) vs. “precision independence” (Yarowsky algorithm)
- Relies on two heuristics and a decision list
 - One sense per collocation :
 - Nearby words provide strong and consistent clues as to the sense of a target word
 - One sense per discourse :
 - The sense of a target word is highly consistent within a single document

Learning Algorithm

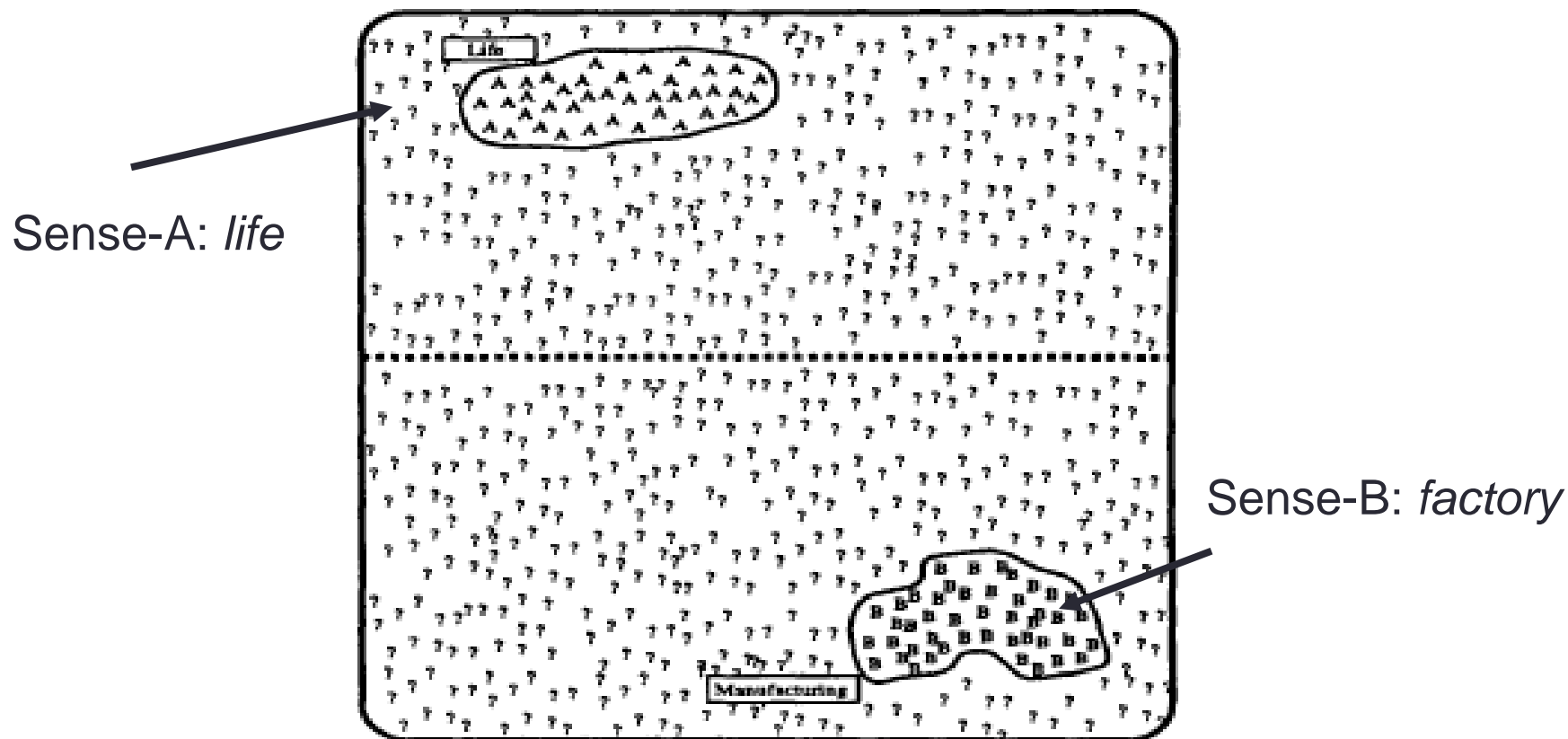
- A decision list is used to classify instances of target word :

“the loss of animal and **plant** species through extinction ...”

- Classification is based on the highest ranking rule that matches the target context

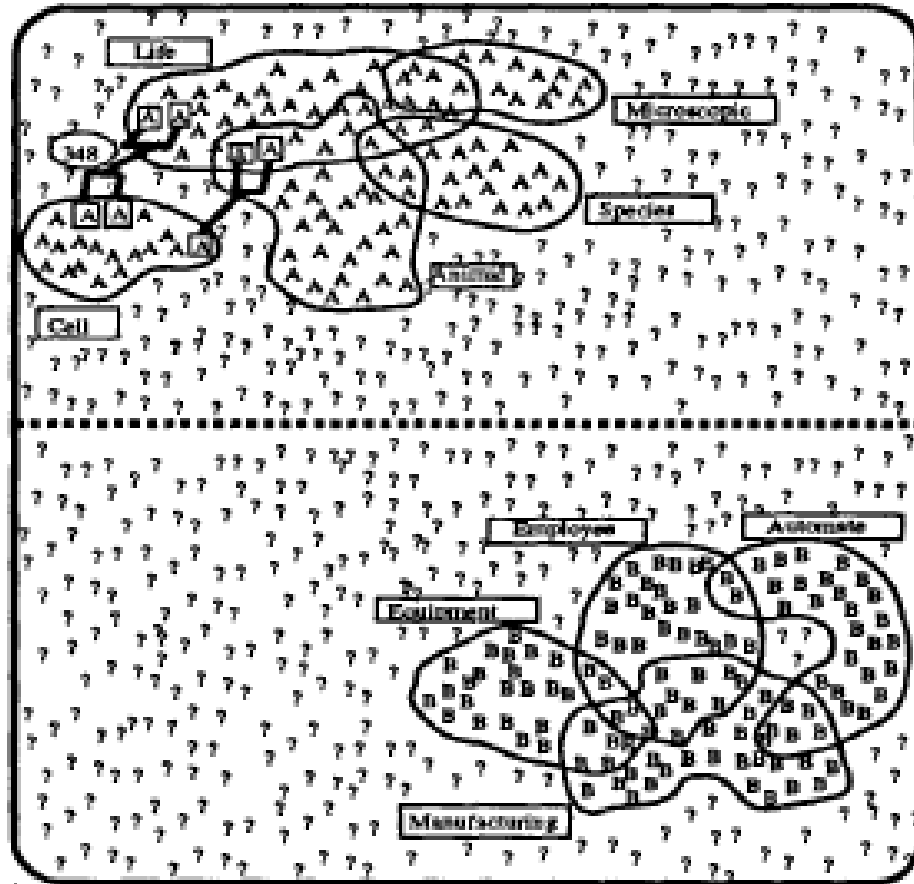
LogL	Collocation	Sense
...
9.31	flower (within +/- k words)	→ A (living)
9.24	job (within +/- k words)	→ B (factory)
9.03	fruit (within +/- k words)	→ A (living)
9.02	<i>plant</i> species	→ A (living)
...

Bootstrapping Algorithm



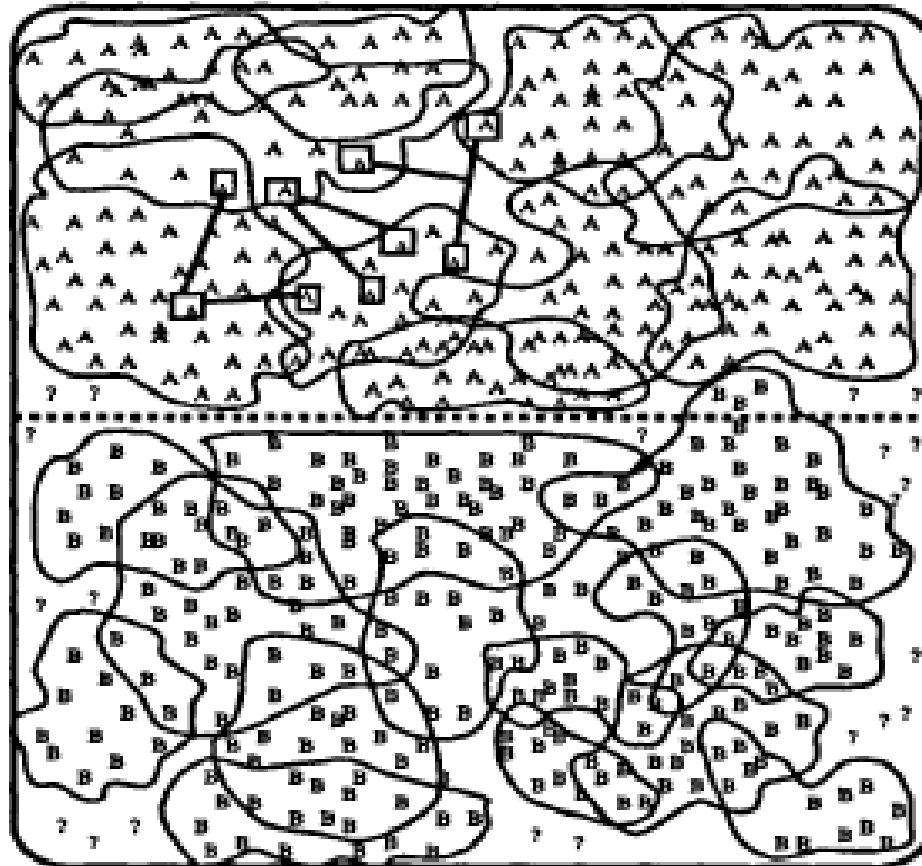
- All occurrences of the target word are identified
- A small training set of seed data is tagged with word sense

Bootstrapping Algorithm



Seed set grows and residual set shrinks

Bootstrapping Algorithm



Convergence: Stop when residual set stabilizes

Bootstrapping Algorithm

- Iterative procedure:
 - Train decision list algorithm on seed set
 - Classify residual data with decision list
 - Create new seed set by identifying samples that are tagged with a probability above a certain threshold
 - Retrain classifier on new seed set
- Selecting training seeds
 - Initial training set should accurately distinguish among possible senses
 - Strategies:
 - Select a single, defining seed collocation for each possible sense.
Ex: “**life**” and “**manufacturing**” for target *plant*
 - Use words from dictionary definitions
 - Hand-label most frequent collocates

Evaluation

- Test corpus: extracted from 460 million word corpus of multiple sources (news articles, transcripts, novels, etc.)
- Performance of multiple models compared with:
 - supervised decision lists
 - unsupervised learning algorithm of Schütze (1992), based on alignment of clusters with word senses

Word	Senses	Supervised	Unsupervised Schütze	Unsupervised Bootstrapping
plant	living/factory	97.7	92	98.6
space	volume/outer	93.9	90	93.6
tank	vehicle/container	97.1	95	96.5
motion	legal/physical	98.0	92	97.9
...	-	...
Avg.	-	96.1	92.2	96.5

References

- (Lesk, 1986) Lesk, M. *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone*. SIGDOC 1986.
- (Yarowsky 1995) Yarowsky, D. *Unsupervised word sense disambiguation rivaling supervised methods*. Proceedings of ACL 1995.
- (Resnik and Yarowsky, 1997) A Perspective on Word Sense Disambiguation Methods and their Evaluation. The ACL-SIGLEX Workshop Tagging Text with Lexical Semantics. pp. 79-86.
- (Schutze, 1998) Automatic Word Sense Discrimination. *Computational Linguistics*, 24 (1) pp. 97-123.
- (Kilgarriff, 1997) “I don’t believe in word senses”, *Computers and the Humanities* (31) pp. 91-113.

Overview

- Machine Learning, Semantics and NLP
(Trattamento Automatico delle Lingue)
 - Objectives,
 - Methods,
 - Resources and Technologies
 - Applications
- Semantics in Language Processing
- Lexical Semantic tasks and Resources
- Predicate Semantics and Role Labeling
- The role of Tree Kernels
- Conclusions

From Lexical to Computational Semantics

Selectional Preferences

- A way to constrain the possible meanings of words in a given context
- E.g. “Wash a dish” vs. “Cook a dish”
 - WASH-OBJECT vs. COOK-FOOD
- Capture information about possible relations between semantic classes
 - Common sense knowledge
- Alternative terminology
 - Selectional Restrictions
 - Selectional Preferences
 - Selectional Constraints

Syntactic Argument Structures

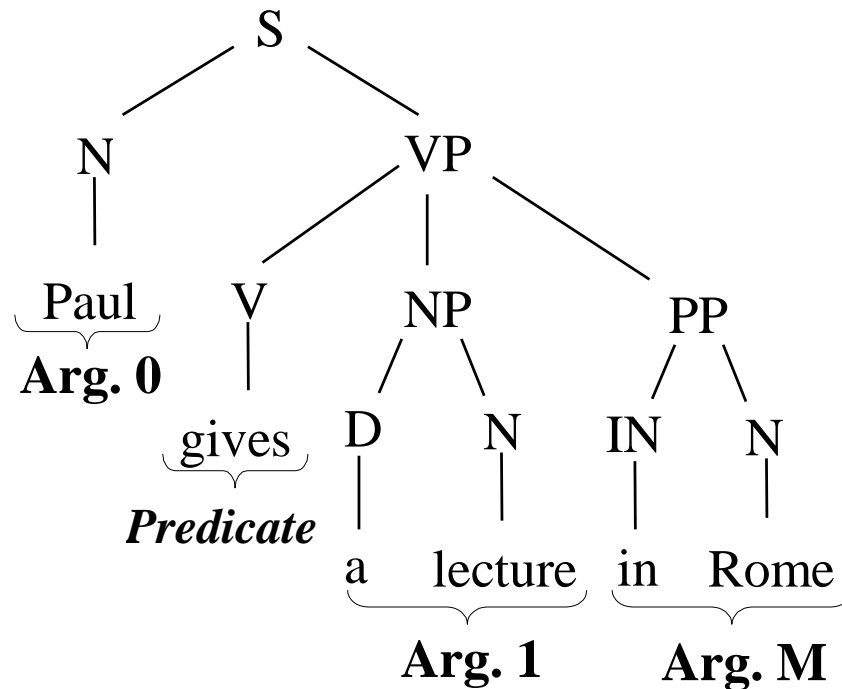
- Compositionality in the treatment of verbs suggests to see them as n -ary relations, partially saturated functions
- (Verbal) Relations determine a fixed number of participants, called **arguments**
- The syntactic structure predicts the number and type of arguments through **subcategorization frames**
 - (Bob (gave (Mary) (the book) (on Monday)))
 - (Bob (gave (the book) (to Mary) (on Monday)))

Thematic roles

- Arguments play specific roles, called **thematic roles**, depending on the predicate but invariant across different syntactic structures giving rise to **predicate argument structures**
 - *give* (Agent: *Bob*, Theme: *the_book*, Recipient: *Mary*)
- Thematic roles of individual arguments are indexed by their predicates
- *General* and *lexicalized* roles have been introduced

Predicate and Arguments

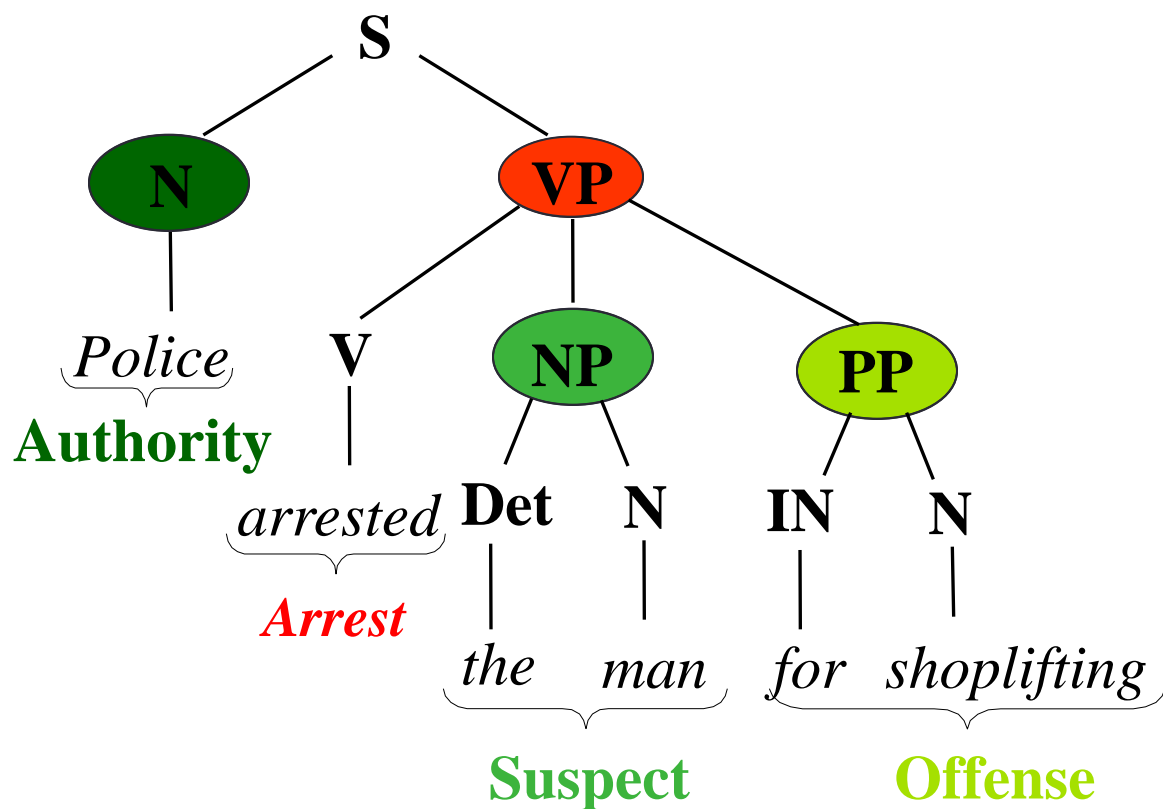
- The syntax-semantic mapping



- Different semantic annotations (e.g. PropBank vs. FrameNet)

Linking syntax to semantics

- *Police arrested the man for shoplifting*



Semantics and News

Applicazioni Risorse Sistema

Gmail ... x SRL_EN x Come ... x R Econo... x Googl... x Tanl It... x Frame... x SRL_EN x Econo... x

file:///home/danilo/Downloads/SRL_ITA/sorgente/Economia%20-%20Repubblica.it.html

Telefilm in stream... Flash Forward pri... Cronologia Altri Pr...



L'ad punta a nuove regole sulla base del modello Pomigliano. L'annuncio, che prevede l'uscita da Federmeccanica, domani al vertice con il governo o giovedì con una lettera a Bombassei. Potrebbe avvenire assieme alla decisione di creare una new company per

Pomigliano di SALVATORE TROPEA

Cisl-Uil: "L'accordo di categoria non si tocca" di S. PAROLA

Saconi: "Su Fiat partita aperta"

Nasce Fabbrica Italia Pomigliano

Si dimette il capo di Bp buonuscita un milione di sterline



Oggi l'annuncio: a Tony Hayward subentrerà il direttore esecutivo Robert Dudley. **I costi legati al disastro sono saliti a 32,2 miliardi di dollari**, ma la società li deterrà evitando di versare al fisco Usa 10 miliardi

Manager Usa, è Ellison di Oracle il più pagato del decennio



Ha guadagnato 1,84 miliardi di dollari. Nella classifica del *Wall Street Journal* sui leader delle società quotate, secondo con 1,14 miliardi il capo di Expedia, terzo Irani di Occidental Petroleum. Solo quarto Steve Jobs

Il nemico alle porte

La Consob e la mano invisibile

Altri articoli

PICCOLE GRANDI IMPRESE

DI LUCA PAGNI

La grande sfida del teleshopping

La crisi colpisce anche i porti turistici ma siamo sicuri che sia un male?

Altri articoli

PERCENTUALMENTE DI ROSARIA AMATO

La prova del 9

L'export risolve il Pil, ma non le famiglie

Altri articoli

GLI ESPERTI RISPONDONO

CASA

A cura di Antonella Donati

Compenso extra, quando ne ha diritto l'amministratore

Mia moglie ed il fratello sono proprietari di un appartamento in condominio. Allo stato

Il tuo libro arriva dove hai sempre sognato.

ilmiolibro.it

24ORE AGI

Roma 19:04
ACEA: NEL I SEMESTRE UTILE NETTO +52,1% A 2010, MLN

Parigi 18:42
AIR FRANCE-KLM: TORNA IN UTILE NEL PRIMO TRIMESTRE

← 3 → Le altre notizie

CREDITO ALLE IMPRESE

Microimprese: con la crisi aumenta il rischio di credito

IN COLLABORAZIONE CON



Semantic [role] LABELING



I costi legati al disastro sono saliti a 32,2 miliardi di dollari.

Submit

Clear

Change_position_on_a_scale: *[I costi]_{Item} legati al disastro sono saliti [a 32,2 miliardi di dollari.]_{Final_value}.*

[Show CONLL format](#)

1	I	RD	det	2	_	Item
2	costi	S	subj	7	_	Item
3	legati	V	mod	2	_	_
4	al	EA	comp	3	_	_
5	disastro	S	prep	4	_	_
6	sono	VA	aux	7	_	_
7	saliti	V	ROOT	0	Change_position_on_a_scale	Target
8	a	E	comp	7	_	Final_value

Completato

- I costi legati al disastro sono saliti a 32,2 miliardi di dollari.



Semantic [role] LABELING

L'aumento dei costi legati al disastro è stato di 32,2 miliardi di dollari.

Submit

Clear

Change_position_on_a_scale: L'aumento [*dei costi*]_{Item} legati al disastro è stato [*di 32,2 miliardi di dollari*]_{Final_value}.

Semantics in NLP: Resources

- Lexicalized Models
 - Propbank
 - NomBank
- Framenet
 - Inspired by *frame semantics*
 - Frames are lexicalized prototypes for real -world situations
 - Participants are called frame elements (roles)

PropBank (Palmer et al., 2005)

- Transfer sentences to propositions
 - *Kristina hit Scott* \Rightarrow hit(Kristina, Scott)
- Penn TreeBank \Rightarrow PropBank
- Add a semantic layer on Penn TreeBank
 - Define a set of semantic roles for each verb
 - Each verb's roles are numbered.

- [A0/the company] to ... offer[A1/a 15% to 20% stake] [A2/to the public].
- [A0/Sotheby's] .offered[A2/the Dorrance heirs] [A1/a money-back
guarantee].
- [A1/an amendment] offered[A0/by Rep. Peter DeFazio] ..
- [A2/Subcontractors] will be offered[A1/a settlement] .

PropBank (2)

- It is difficult to define a general set of semantic roles for all types of predicates (verbs).
- PropBank defines semantic roles for each verb and sense in the frame files.
- The (core) arguments are labeled by numbers.
 - A0 -Agent; A1 -Patient or Theme
 - Other arguments -no consistent generalizations
- Adjunct-like arguments -universal to all verbs
 - AM-LOC, TMP, EXT, CAU, DIR, PNC, ADV, MNR, NEG, MOD, DIS

PropBank (3) – an example

Predicate *extract*:

Frames file for 'extract' based on survey of sentences in the WSJ corpus.

Roleset extract.01 **Verbnet Class: 1** "to remove or obtain":

Roles:

Arg0: remover, agent

Arg1: thing extracted

Arg2: entity extracted from

Examples:

person: *ns* tense: *ns* aspect: *ns* voice: *active* form: *infinitive*

transitive (-)

[Bond investors]-2 paid close attention to comments by Federal Reserve Chairman Alan Greenspan , who [*T*-1] was testifying be hearing , but were n't able [*-2] to extract many clues about the future course of the Fed 's monetary policy .

Arg0: [*-2]

REL: extract

Arg1: many clues about the future course of the Fed 's monetary policy

person: *ns* tense: *ns* aspect: *ns* voice: *active* form: *infinitive*

all arguments (-)

PropBank – Frame files

- hit.01 “strike”

- ❖ A0: agent, hitter; A1: thing hit;
A2: instrument, thing hit by or with

[_{A0} *Kristina*] **hit** [_{A1} *Scott*] [_{A2} *with a baseball*] *yesterday*.

AM-TMP
Time

- look.02 “seeming”

- ❖ A0: seemer; A1: seemed like; A2: seemed to

[_{A0} *It*] **looked** [_{A2} *to her*] *like* [_{A1} *he deserved this*].

- deserve.01 “deserve”

- ❖ A0: deserving entity; A1: thing deserved;
A2: in-exchange-for

It looked to her like [_{A0} *he*] **deserved** [_{A1} *this*].

Proposition:
A sentence and
a target verb

PropBank – Data

- as for release by Mar 4, 2005
- **Proposition Bank I**
 - Verb Lexicon: 3,324 frame files
 - Annotation: ~113,000 propositions
 - http://www.cis.upenn.edu/~mpalmer/project_pages/ACE.htm
- Alternative format: CoNLL-04,05 shared task
 - Represented in table format
 - Has been used as standard data set for the shared tasks on semantic role labeling
 - <http://www.lsi.upc.es/~srlconll/soft.html>

Frame Semantics

- Research in Empirical Semantics suggests that words represents categories of experience (*situations*)
- A *frame* is a cognitive structuring device (i.e. a kind of prototype) indexed by *words* and used to support understanding (Fillmore, 1975)
 - Lexical Units evoke a Frame in a sentence
- Frames are made of *elements* that express participants to the situation (Frame Elements)
- During communication LUs evoke the frames

Frame Semantics

Frame: KILLING	
A KILLER or CAUSE causes the death of the VICTIM.	
Frame Elements	KILLER John <u>drowned</u> Martha.
	VICTIM John <u>drowned</u> Martha .
	MEANS The flood <u>exterminated</u> the rats by cutting off access to food .
	CAUSE The rockslide <u>killed</u> nearly half of the climbers.
	INSTRUMENT It's difficult to <u>suicide</u> with only a pocketknife .
Predicates	annihilate.v, annihilation.n, asphyxiate.v, assassin.n, assassinate.v, assassination.n, behead.v, beheading.n, blood-bath.n, butcher.v, butchery.n, carnage.n, crucifixion.n, crucify.v, deadly.a, decapitate.v, decapitation.n, destroy.v, dispatch.v, drown.v, eliminate.v, euthanasia.n, euthanize.v, ...

Frame Semantics

- Lexical descriptions are expected to define the indexed frame and the frame elements with their realization at the syntactic level:
 - *John bought a computer from Janice for 1000 \$*
- Mapping into syntactic arguments
 - the buyer is (usually) in the subject position
- Obligatory vs. optional arguments
- Selectional preferences
 - *The seller* and *the buyer* are usually “humans” or “social groups”

The FrameNet project

- The aims
 - Create a lexical resource by describing a significant portion of English in terms of precise and rich frame semantics
- The output
 - Frame Database: a structured system of Frames and Fes
 - Lexical database: syntactic and semantic descriptions of frame-evoking words (N,V,A)
 - Annotated Corpus: wide coverage examples



Frame Report (recent data)

[Top of Frame Index](#) | [Top of Lexical Unit Index](#)

Committing_crime

Definition:

A **Perpetrator** (generally intentionally) commits a **Crime**, i.e. does something not permitted by the laws of society.

They PERPETRATED a felony by substituting a lie for negotiations.

The suspect had allegedly COMMITTED the crime to gain the attention of a female celebrity.

FEs:

Core:

Crime [Cr]

An act, generally intentional, that has been formally forbidden by law.

How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Perpetrator [Perp]

The individual that commits a **Crime**.

How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Non-Core:

Frequency [Freq]

The frequency with which a **Crime** is committed.

The average serial killer COMMITTS a crime every five years.







Instrument [Inst]

The **Instrument** used in committing the crime.

Most crimes are COMMITTED with a firearm.

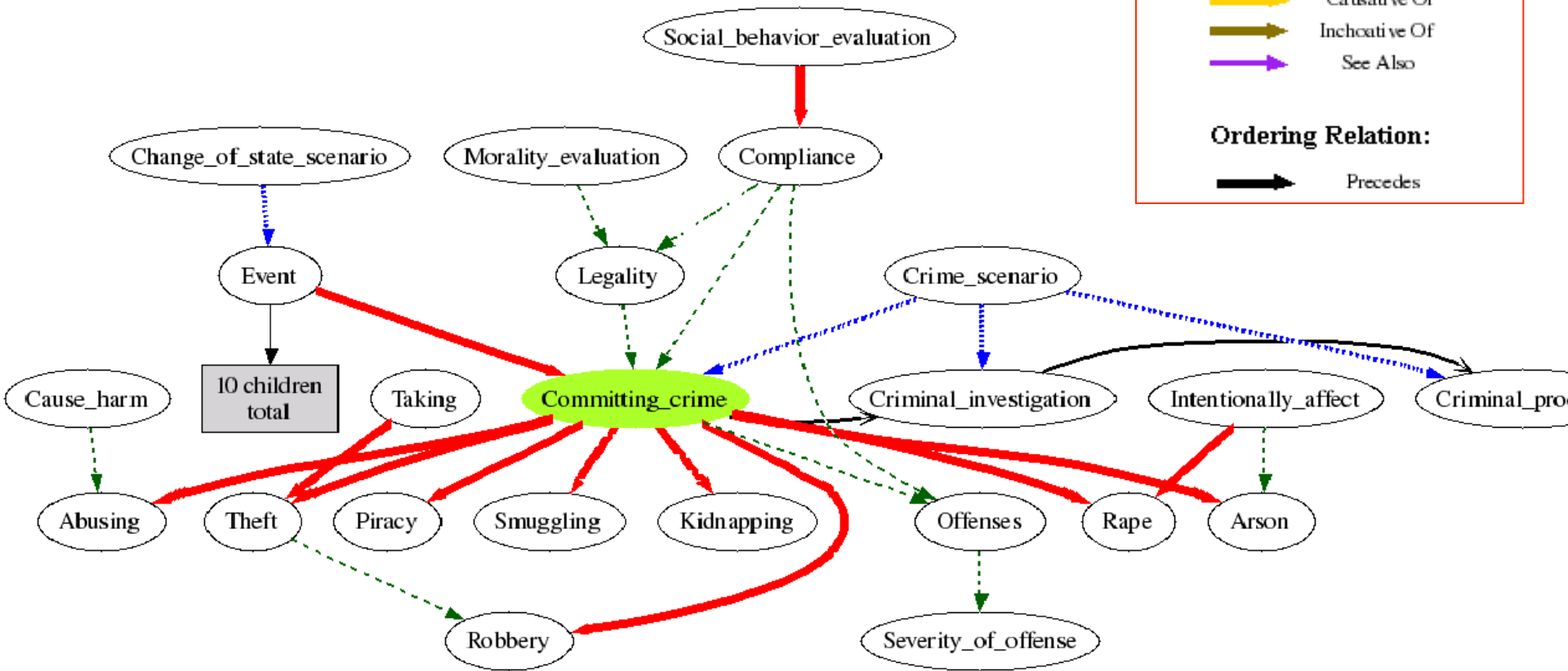
Parent frame → Child frame

Parent → Child Relation Types:

-  Inheritance
-  Subframe
-  Perspective On
-  Using
-  Causative Of
-  Indicative Of
-  See Also

Ordering Relation:

-  Precedes



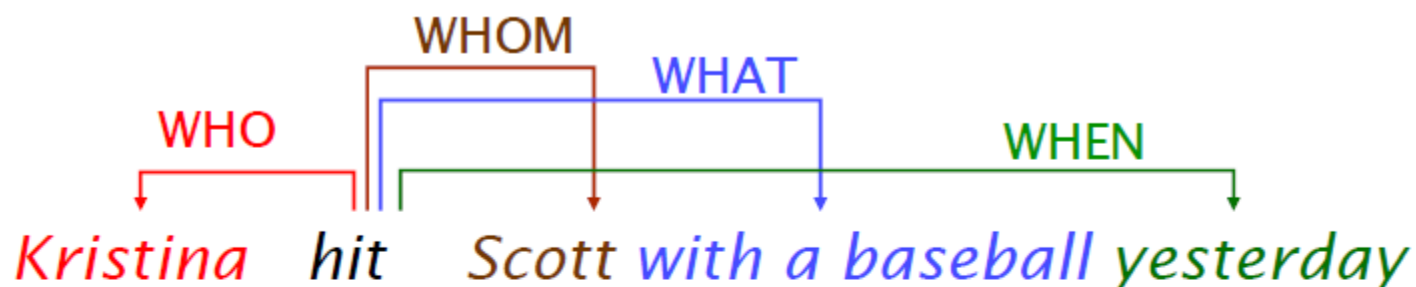
FrameNet - Data

- Methodology of constructing FrameNet
 - Define/discover/describe frames
 - Decide the participants (frame elements)
 - List lexical units that evoke the frame
 - Find example sentences in the BNC and annotate them
- Corpora
 - FrameNet I -British National Corpus only
 - FrameNet II -LDC North American Newswire corpora
- Size
 - >10,000 lexical units, >825 frames, >135,000 sentences
- <http://framenet.icsi.berkeley.edu>



Recognizing Predicates: SRL

- Semantic role labeling vs. QA



- **Who** hit Scott with a baseball?
- **Whom** did Kristina hit with a baseball?
- **What** did Kristina hit Scott with?
- **When** did Kristina hit Scott with a baseball?

Roles and variants in QA

Yesterday, Kristina hit Scott with a baseball

Scott was hit by Kristina yesterday with a baseball

Yesterday, Scott was hit with a baseball by Kristina

With a baseball, Kristina hit Scott yesterday

Yesterday Scott was hit by Kristina with a baseball

Kristina hit Scott with a baseball yesterday

Agent, hitter

Thing hit

Instrument

Temporal adjunct

SRL: task formulation

- Most general formulation: determine a labeling on (usually but not always contiguous) *substrings* (*phrases*) of the sentence s , given a predicate p

$[_{A_0}$ The queen] **broke** $[_{A_1}$ the window].

$[_{A_1}$ By working hard], $[_{A_0}$ he] **said**, $[_{C-A_1}$ you can get exhausted].

- Every substring c can be represented by a set of word indices $c \subseteq \{1, 2, \dots, m\}$
- More formally, a semantic role labeling is a mapping from the set of substrings of s to the label set \mathbf{L} . \mathbf{L} includes all argument labels and NONE.

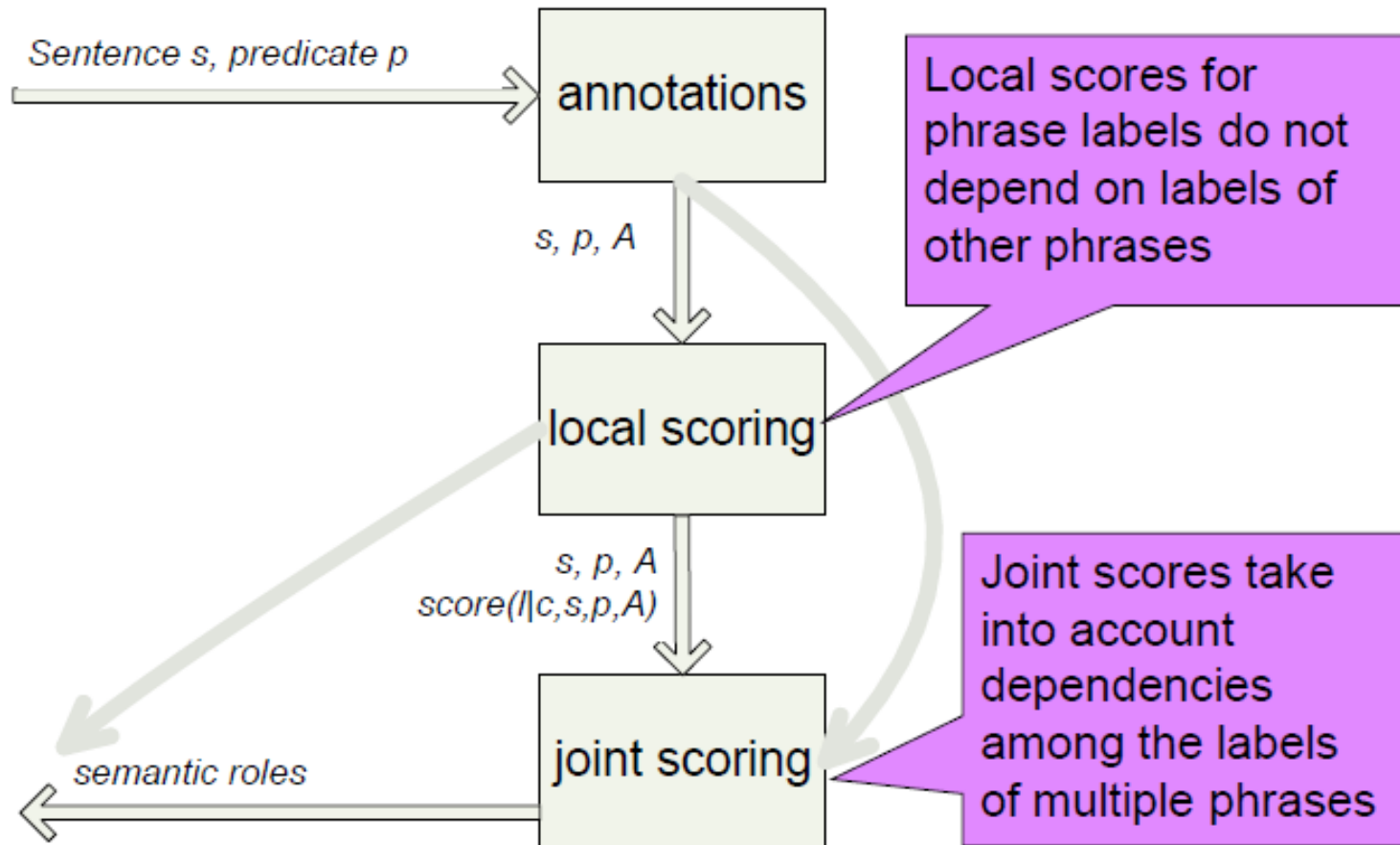
The SRL cascade

- Identification:
 - Very hard task: to separate the argument substrings from the rest in this exponentially sized set
 - Usually only 1 to 9 (avg. **2.7**) substrings have labels ARG and the rest have NONE for a predicate
- Classification:
 - Given the set of substrings that have an ARG label, decide the exact semantic label
- Core argument semantic role labeling: (easier)
 - Label phrases with core argument labels only. The modifier arguments are assumed to have label NONE.

ML Approaches

- **Local models** decide the label of each substring independently of the labels of other substrings
- This can lead to inconsistencies
 - overlapping argument strings
*By [_{A1} working [_{A1} hard] , he] **said** , you can achieve a lot.*
 - repeated arguments
*By [_{A1} working] hard , [_{A1} he] **said** , you can achieve a lot.*
 - missing arguments
*[_{A0} By working hard , he] **said** , [_{A0} you can achieve a lot].*
- **Joint models** take into account the dependencies among labels of different substrings

The general SRL architecture



Previous work on Local ...

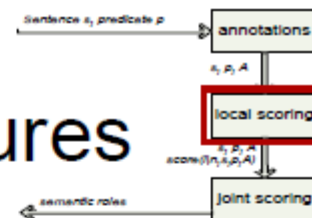
- [Gildea&Jurafsky 02]
 - **Identification + Classification** for local scoring experiments
 - **One Step** for joint scoring experiments
- [Xue&Palmer 04] and [Punyakanok et al. 04, 05]
 - **Pruning + Identification + Classification**
- [Pradhan et al. 04] and [Toutanova et al. 05]
 - **One Step**

... and Joint SRL models

- Tight integration of local and joint scoring in a single probabilistic model and exact search [Cohn&Blunsom 05] [Màrquez et al. 05],[Thompson et al. 03]
 - When the joint model makes strong independence assumptions
- Re-ranking or approximate search to find the labeling which maximizes a combination of local and a joint score [Gildea&Jurafsky 02] [Pradhan et al. 04] [Toutanova et al. 05] [Moschitti et al. 07]
 - Usually exponential search required to find the exact maximizer
- Exact search for best assignment by local model satisfying hard joint constraints
 - Using Integer Linear Programming [Punyakanok et al 04,05] (worst case NP-hard)

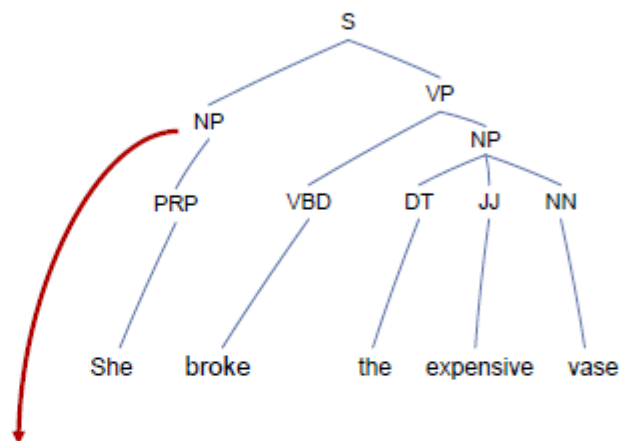
Features (for Local models)

Gildea & Jurafsky (2002) Features



- Key early work
 - Future systems use these features as a baseline

- Constituent Independent
 - Target predicate (lemma)
 - Voice
 - Subcategorization
- Constituent Specific
 - Path
 - Position (*left, right*)
 - Phrase Type
 - Governing Category (*S* or *VP*)
 - Head Word



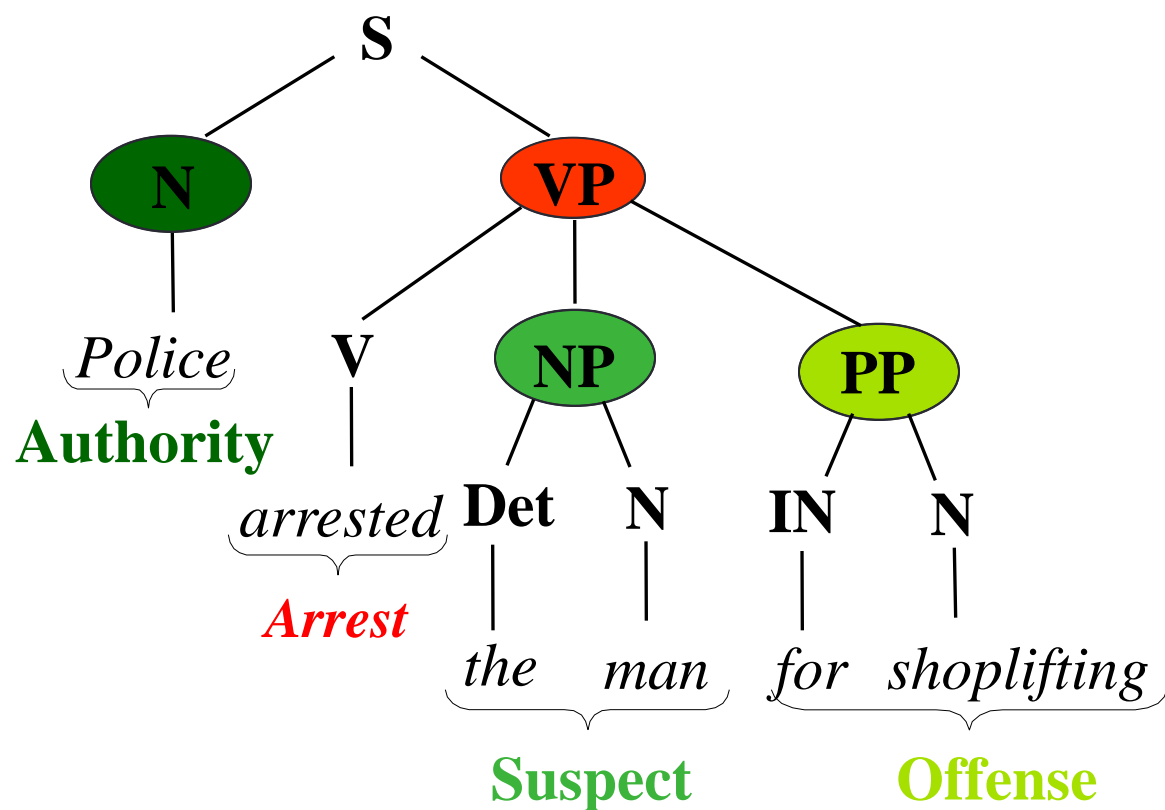
Target	<i>broke</i>
Voice	<i>active</i>
Subcategorization	<i>VP → VBD NP</i>
Path	<i>VBD ↑ VP ↑ S ↓ NP</i>
Position	<i>left</i>
Phrase Type	<i>NP</i>
Gov Cat	<i>S</i>
Head Word	<i>She</i>

Overview

- Machine Learning, Semantics and NLP
(Trattamento Automatico delle Lingue)
 - Objectives,
 - Methods,
 - Resources and Technologies
 - Applications
- Semantics in Language Processing
- Lexical Semantic tasks and Resources
- Predicate Semantics and Role Labeling
- The role of Tree Kernels
- Conclusions

Semantic Role Labeling with TKs

- *Police arrested the man for shoplifting*



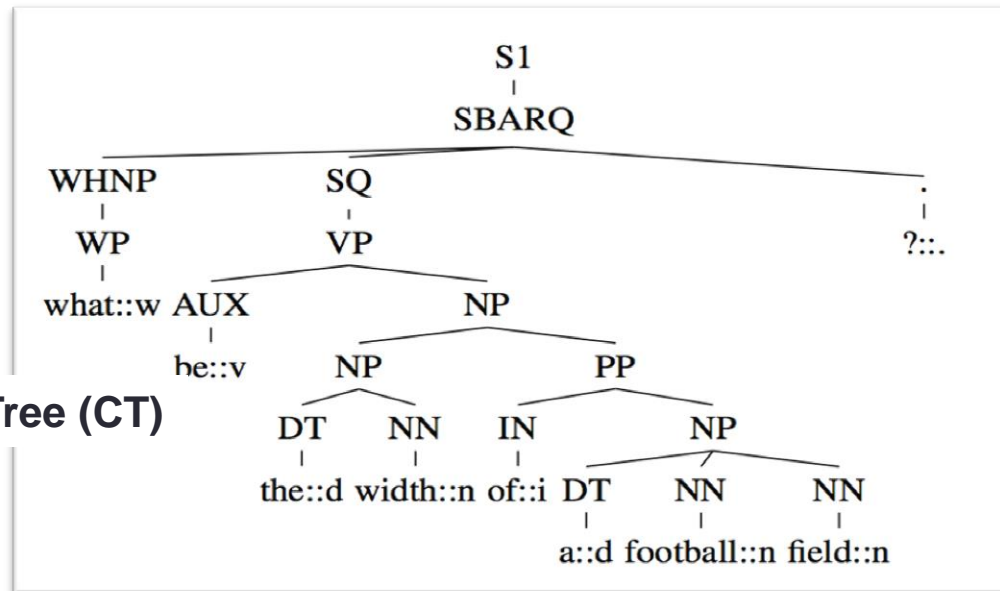
Motivations

- **Limitations of SRL systems:**
- *Manual features cannot apply to specific relational tasks*
- *Poor lexical generalization*
- *Annotation costs*
- *Risk of overfitting*

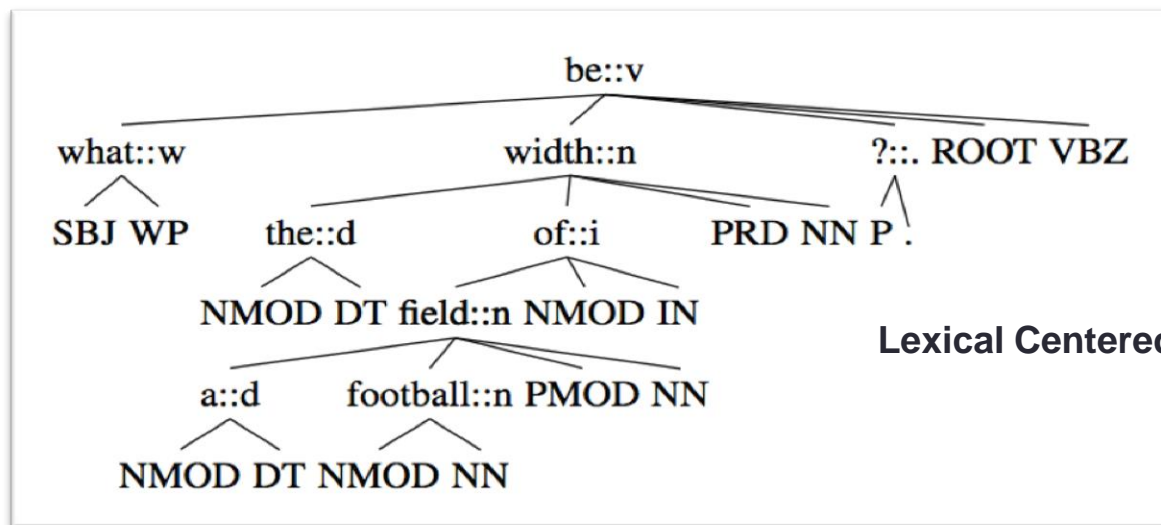
Goals

- **Goals:**
- Develop a semi-supervised statistical SRL model exploiting distributional analysis of unlabeled corpora. LSA embedding is applied to labeled examples.
- Avoid data overfitting using a simple feature set including:
 - the grammatical relation r between the predicate and the argument head

Form of the source Grammatical Trees

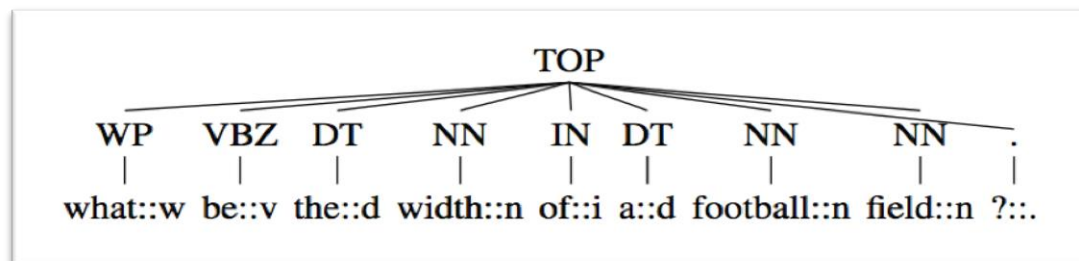
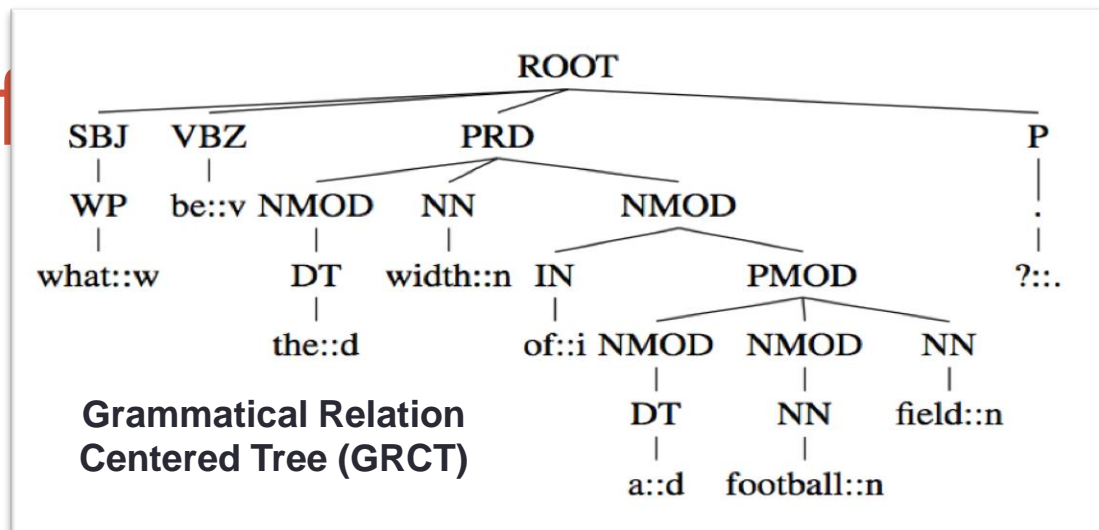


Constituent Tree (CT)

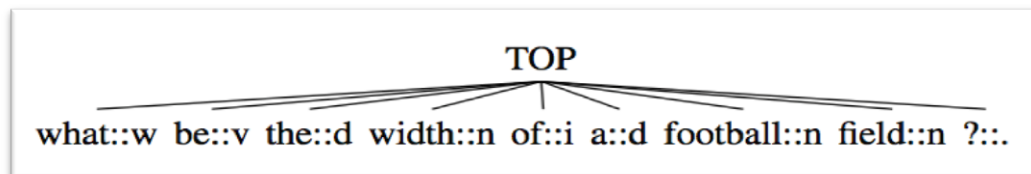


Lexical Centered Tree (LCT)

Form of



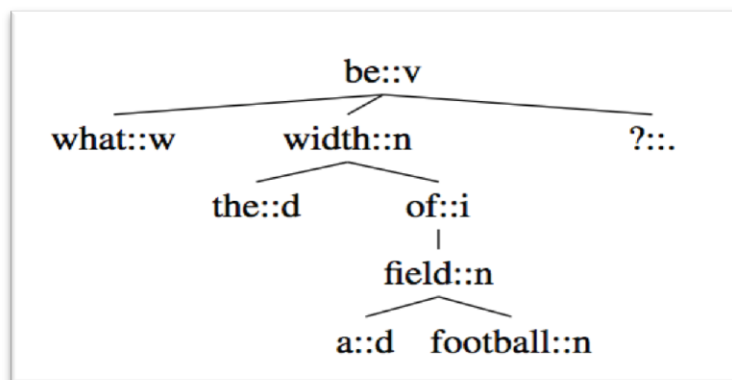
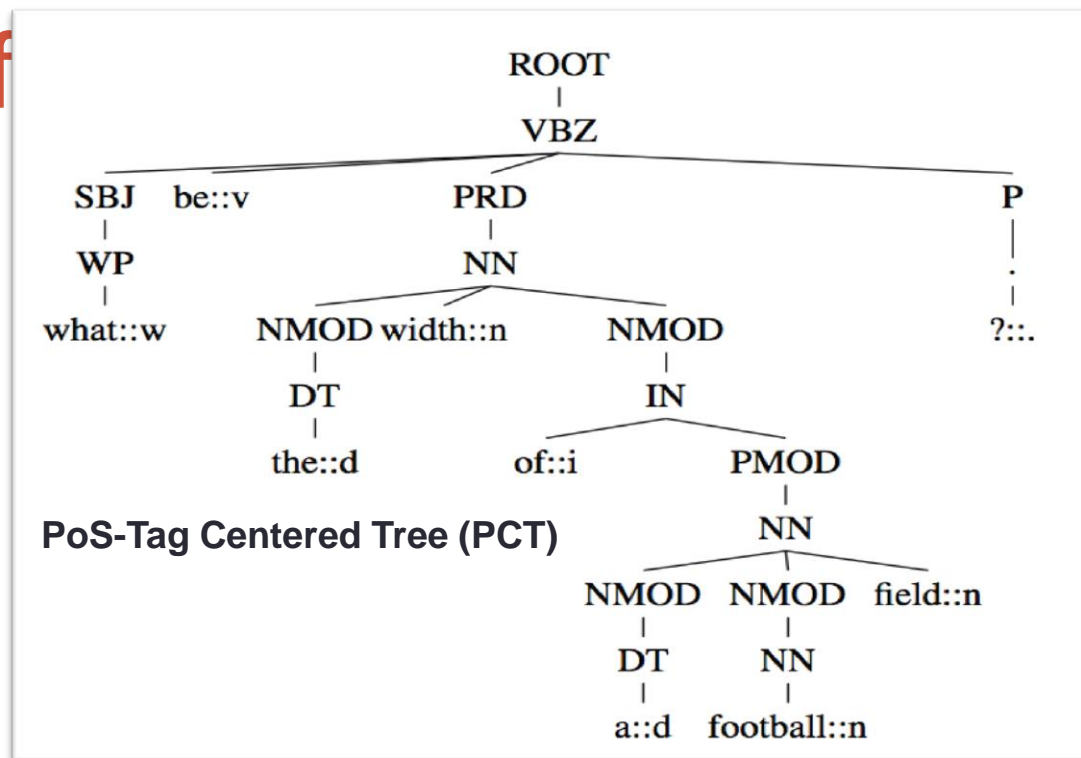
Lexical and PoS-Tag Sequences Tree (LPST)



Lexical Sequences Tree (LST)

Form of

S



Lexical Only Centered Tree (LOCT)

Tree kernels formulation

- PTKs

$$\Delta_{PTK}(n_1, n_2) = \mu \left(\lambda^2 + \sum_{\vec{I}_1, \vec{I}_2, l(\vec{I}_1)=l(\vec{I}_2)} \lambda^{d(\vec{I}_1)+d(\vec{I}_2)} \prod_{j=1}^{l(\vec{I}_1)} \Delta_{PTK}(c_{n_1}(\vec{I}_{1j}), c_{n_2}(\vec{I}_{2j})) \right)$$

- Smoothed Partial Tree Kernel

If n_1 and n_2 are leaves then $\Delta_\sigma(n_1, n_2) = \mu\lambda\sigma(n_1, n_2)$; else

$$\Delta_\sigma(n_1, n_2) = \mu\sigma(n_1, n_2) \times \left(\lambda^2 + \sum_{\vec{I}_1, \vec{I}_2, l(\vec{I}_1)=l(\vec{I}_2)} \lambda^{d(\vec{I}_1)+d(\vec{I}_2)} \prod_{j=1}^{l(\vec{I}_1)} \Delta_\sigma(c_{n_1}(\vec{I}_{1j}), c_{n_2}(\vec{I}_{2j})) \right), \quad (2)$$

A small example:

- How can we estimate the similarity between:
 “man reads magazine” and *“woman browses newspaper”*
- In (Clark&Pulman07) a tensor based operator has been proposed



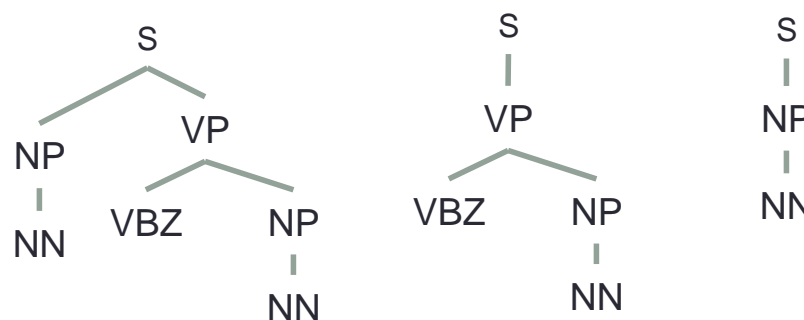
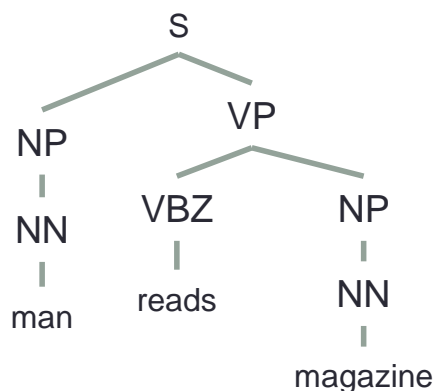
$(man \ddot{\wedge} reads \ddot{\wedge} magazine) \times (woman \ddot{\wedge} browses \ddot{\wedge} newspaper)$

$$(w_1 \ddot{\wedge} w_2) \times (w_3 \ddot{\wedge} w_4) = (w_1 \times w_3) \dot{\wedge} (w_2 \times w_4)$$

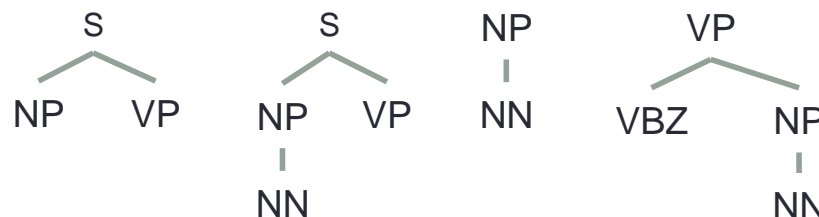
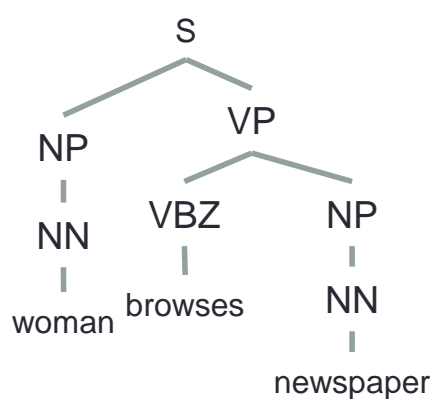
$(man \times woman) \dot{\wedge} (reads \times browses) \dot{\wedge} (magazine \times newspaper)$

The role of Partial Tree Kernels

We count the common subtrees



...



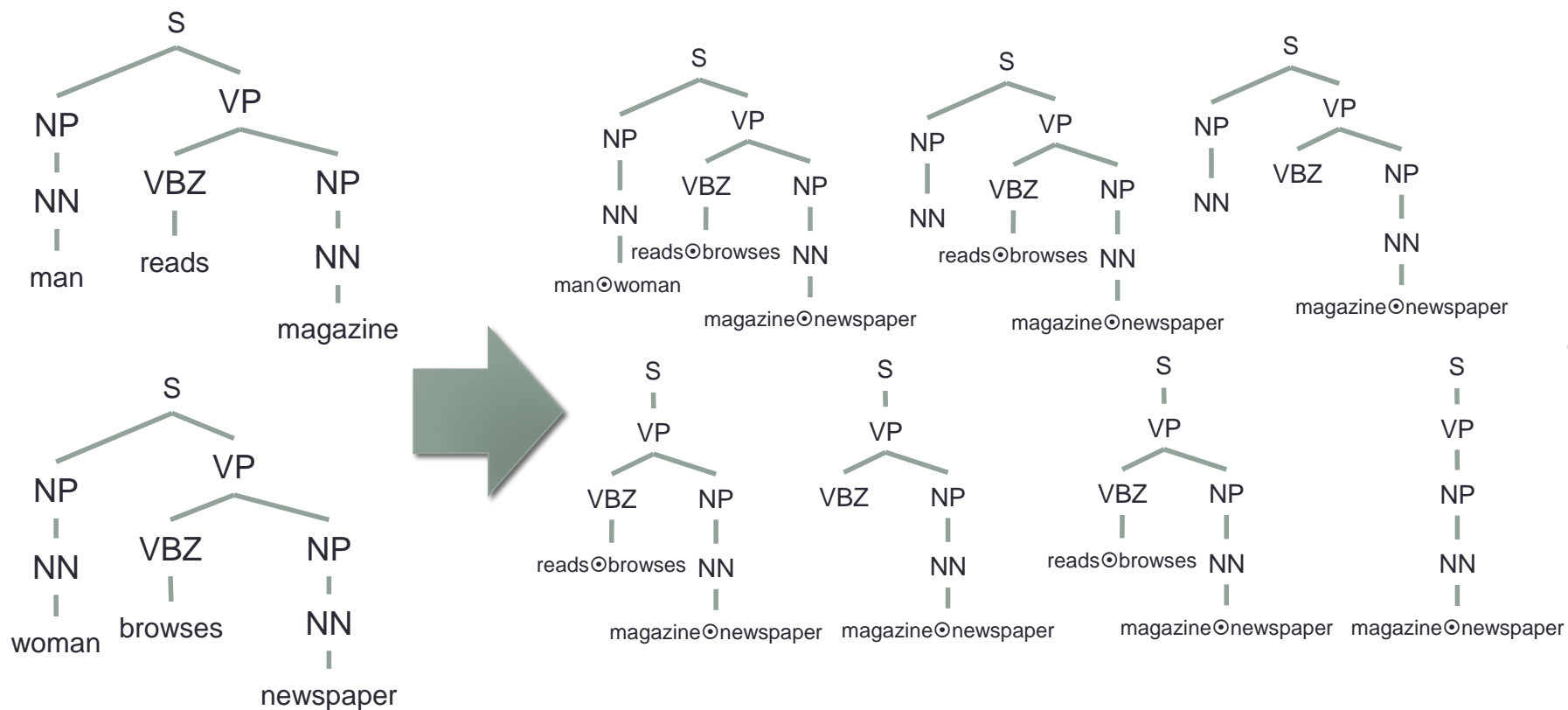
Each subtrees corresponds to a feature in a hidimensional space that is not explicitly computed

What about lexical information?

What about the sentence "dogs bite man" ?

The role of lexical information

We can treat equivalently words like “*man*⊙*woman*”



What kind of feature space is generated?

Smoothed Partial Tree Kernel

- If n_1 and n_2 are leaves then

$$\Delta_\sigma(n_1, n_2) = \mu\lambda\sigma(n_1, n_2)$$

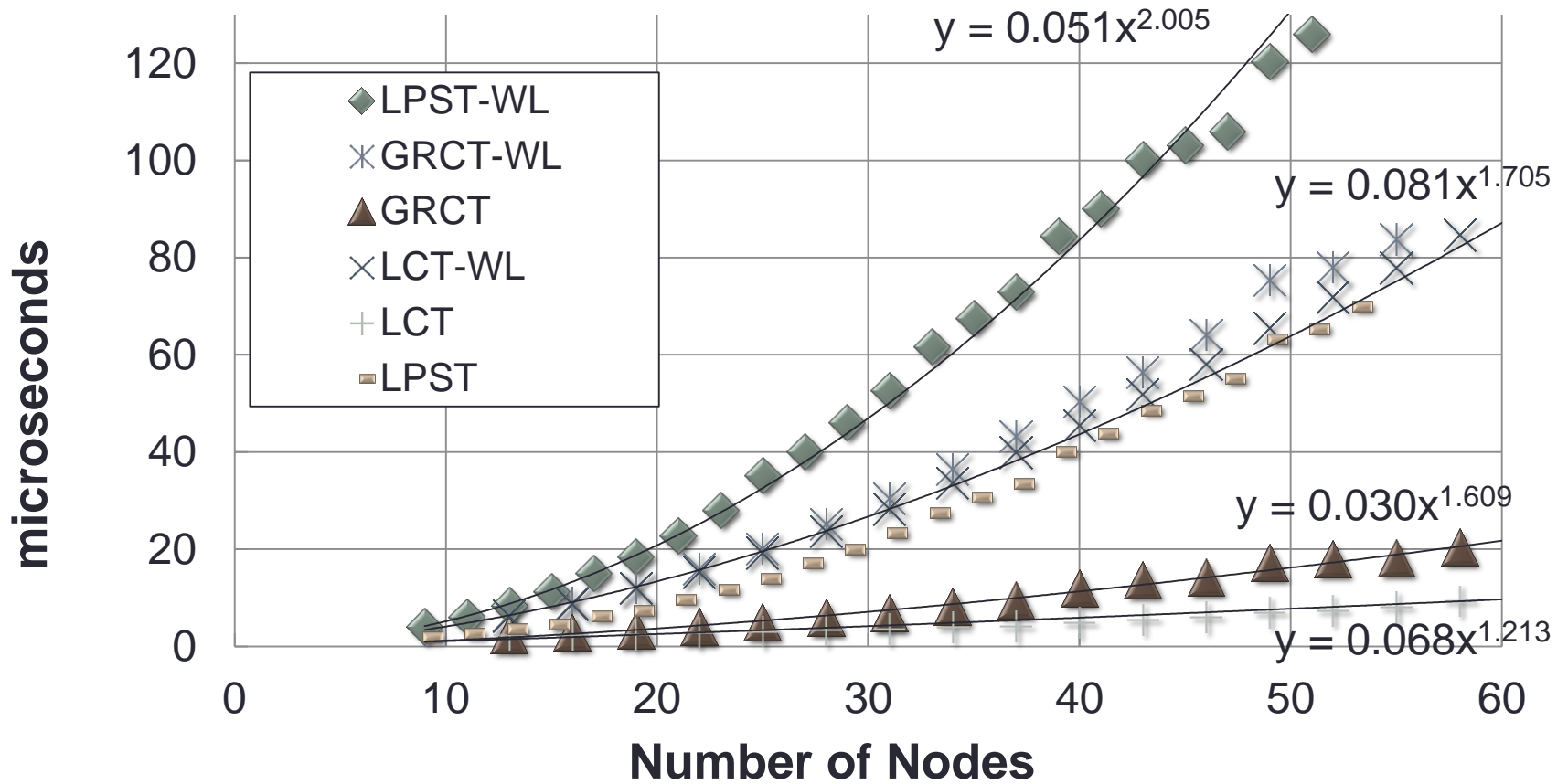
else

$$\Delta_\sigma(n_1, n_2) = \mu\sigma(n_1, n_2) \times \left(\lambda^2 + \sum_{\vec{I}_1, \vec{I}_2, l(\vec{I}_1)=l(\vec{I}_2)} \lambda^{d(\vec{I}_1)+d(\vec{I}_2)} \prod_{j=1}^{l(\vec{I}_1)} \Delta_\sigma(c_{n_1}(\vec{I}_{1j}), c_{n_2}(\vec{I}_{2j})) \right)$$

Experimental Evaluation

- Two classification tasks: Question Classification and Argument Classification
 - We extended the SVM-LightTK software to implement the SPTK
 - Parameterization of classifiers is carried on a held-out set (30% of the training)
 - We experiment with multi-classification, which we model through one-vs-all.
 - The quality of such classification is measured with accuracy.
- Parser: we used the Charniak parser for generating constituency trees, LTH parser (Johansson and Nugues, 2008) to generate dependency trees.
 - POS tags are used to estimate similarity among words that have the same POS

Computational Performance

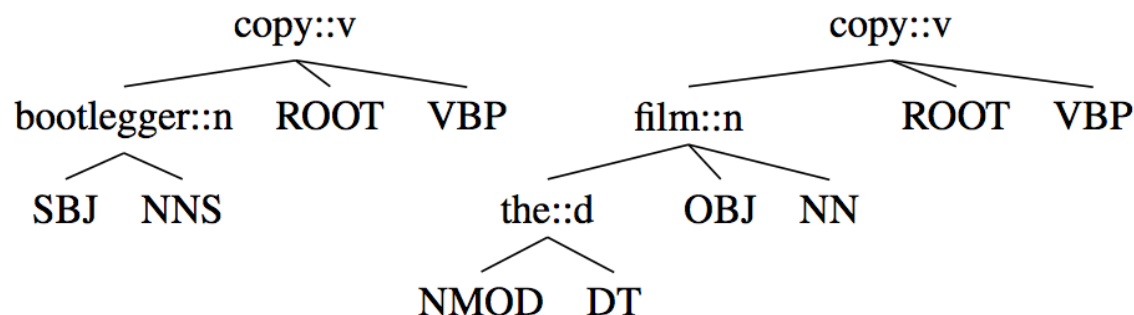


Micro-seconds for each kernel computation

Argument Classification

- We experimented with a FrameNet SRL classification (gold standard boundaries)
- We used the FrameNet version 1.3: 648 frames are considered
 - Test set: 3Training set: 271,560 arguments (90%)
 - 0,173 arguments (10%)

*[Bootleggers]_{CREATOR}, then **copy** [the film]_{ORIGINAL} [onto hundreds of VHS tapes]_{GOAL}*



Kernel	Accuracy
GRCT	87,60%
GRCT _{LSA}	88,61%
LCT	87,61%
LCT _{LSA}	88,74%
GRCT+LCT	87,99%
GRCT _{LSA} +LCT _{LSA}	88,91%

Lexical Similarity

- **Latent Semantic Analysis (LSA)**

- a co-occurrence space is built from ukWak, a document collection made by 2 billion tokens
- the contexts are short windows of size $[-3, +3]$.
- the most frequent 20,000 items are selected along with their 20k contexts.
- the entries of M are the point-wise mutual information between them.
- the SVD reduction is then applied to M , with a dimensionality cut of $l = 250$.

- **Word List (WL):**

- QC task uses also the similarity based on word list provided in (Li and Roth, 2002)
- It is more precise and manually checked

Question Classification

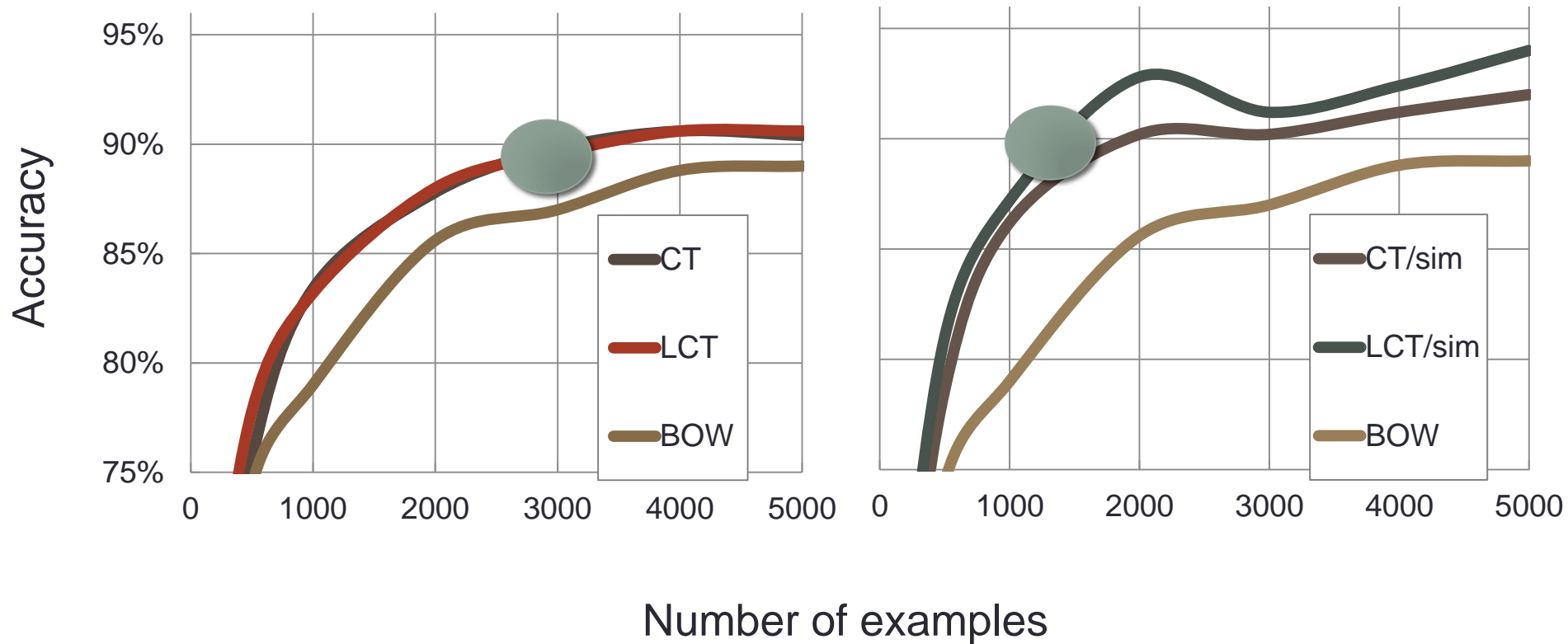
- We used the UIUC dataset (Li and Roth, 2002)
- Question classes are organized in two levels:
 - 6 coarse-grained classes (like ENTITY or HUMAN)
 - 50 fine-grained sub-classes (e.g. Plant, Food as subclasses of ENTITY)
- Training set: 5,452 questions
- Test set: of 500 questions

Kernel	COARSE			FINE		
	NO	LSA	WL	NO	LSA	WL
CT	90,80%	91,00%	92,20%	84,00%	83,00%	86,60%
GRCT	91,60%	92,60%	94,20%	83,80%	83,20%	85,00%
LCT	90,80%	94,80%	94,20%	85,40%	86,20%	87,40%
LOCT	89,20%	93,20%	91,80%	85,40%	86,80%	87,00%
LST	88,20%	85,80%	89,60%	84,00%	80,00%	85,00%
LPST	89,40%	89,60%	92,40%	84,20%	82,20%	84,60%
PCT	91,20%	92,20%	93,40%	84,80%	84,00%	85,20%
CT-STK	91,20%	-	-	82,20%	-	-
BOW	88,80%	-	-	83,20%	-	-

Learning Curve

NO Similarity

Similarity based



Perspectives

- How to build significant lexical resources through ML over corpora
 - WordSpaces over domain collections
 - General purpose lexicons over Wikipedia
 - Use of Social Media data (Facebook, Twitter, ...)
- More flexible supervised learning
 - New metrics (e.g. extensions of tree kernels)
 - Data-driven metrics (e.g. manifold learning, autoencoders)
 - Semi-supervised technologies applied to large scale data sets, through on-line learning systems
- Newer problems:
 - From semantics to emotions, engagement
 - Opinions, recommending and trends: social dynamic phenomena

Conclusions

- Natural language is the main carrier of semantic information across
 - People
 - Media
 - Communities
 - Borders
- Language Processing in Italy is enough mature for its integration within a number of semantic workflows
- Multimedia applications on the Web are also more demanding for coherent and semantically rich annotations

Conclusions (2)

- Machine Learning allows to improve the
 - Quality (accuracy, naturality)
 - Effectiveness/Robustness
 - Scale
 - Efficiency (lower costs)
- ... of language processing systems
- Advanced ML has been applied and inspired by several tasks (e.g. text classification, parsing and semantic role labeling)
- The AI Lab at Tor Vergata a large experience in the development of complex ML-based NLP systems for a variety of semantic inference tasks in Web IR