Semantic Role Labeling
Past, Present and Future

Lluís Màrquez
TALP Research Center
Tecnhical University of Catalonia

Tutorial at ACL-IJCNLP 2009
Suntec – Singapore
August 2, 2009

—Version from August 3, 2009—

1. Introduction
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
5. Conclusions

Tutorial Overview

1 Introduction
- Problem definition and properties
- Main Computational Resources and Systems

2 State-of-the-art

3 Empirical evaluation and lessons learned

4 Problems and challenges

5 Conclusions

Tutorial Overview

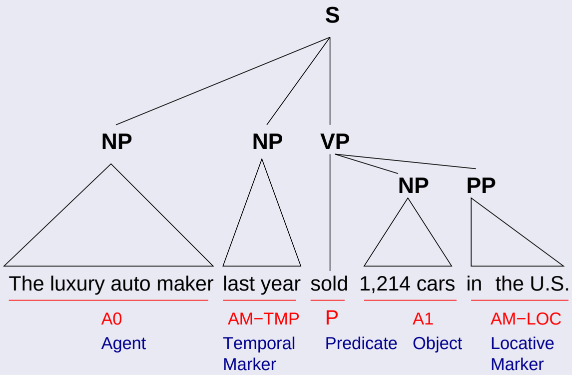
1. Introduction
   - Problem definition and properties
   - Main Computational Resources and Systems
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
5. Conclusions

Semantic Role Labeling: The Problem

SRL $\stackrel{def}{=}$ detecting basic event structures such as *who* did *what* to *whom*, *when* and *where* [**IE** point of view]

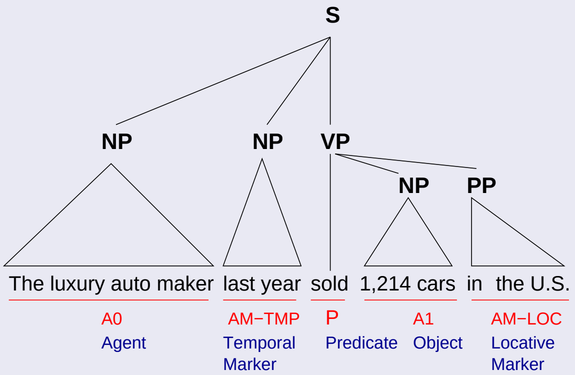
Semantic Role Labeling: The Problem

SRL $\stackrel{def}{=}$ detecting basic event structures such as *who* did *what* to *whom*, *when* and *where* [**IE** point of view]



Semantic Role Labeling: The Problem

SRL $\stackrel{def}{=}$ identify the *arguments* of a given verb and assign them semantic labels describing the *roles* they play in the predicate (i.e., identify predicate argument structures) [**CL** point of view]



Semantic Role Labeling: The Problem

Syntactic variations

$\overbrace{\text{Yesterday}}^{\text{TEMP}}$, $\overbrace{\text{Kristina}}^{\text{HITTER}}$ hit $\overbrace{\text{Scott}}^{\text{THING HIT}}$ $\overbrace{\text{with a baseball}}^{\text{INSTRUMENT}}$

- 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

Example from (Yih & Toutanova, 2006)

Semantic Role Labeling: The Problem

Syntactic variations

$$\overbrace{\text{Yesterday}}^{\text{TEMP}},\ \overbrace{\text{Kristina}}^{\text{HITTER}}\ \text{hit}\ \overbrace{\text{Scott}}^{\text{THING HIT}}\ \overbrace{\text{with a baseball}}^{\text{INSTRUMENT}}$$

- 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

Example from (Yih & Toutanova, 2006)

Semantic Role Labeling: The Problem

Structural view

Mapping from input to output structures:

- **Input** is *text* (enriched with morpho-syntactic information)
- **Output** is a *sequence of labeled arguments*
- **Sequential** segmenting/labeling problem

" Mr. Smith sent the report to me this morning . "

[Mr. Smith]$_{AGENT}$ sent [the report]$_{OBJ}$ to [me]$_{RECIP}$ [this morning]$_{TMP}$.

Mr.$_{B-AGENT}$ Smith$_{I}$ sent the$_{B-OBJ}$ report$_{I}$ to$_{O}$ me$_{B-RECIP}$ this$_{B-TMP}$ morning$_{I}$.$_{O}$

Semantic Role Labeling: The Problem

Structural view

Mapping from input to output structures:

- **Input** is *text* (enriched with morpho-syntactic information)
- **Output** is a *sequence of labeled arguments*
- **Sequential** segmenting/labeling problem

" Mr. Smith sent the report to me this morning . "

[Mr. Smith]$_{AGENT}$ sent [the report]$_{OBJ}$ to [me]$_{RECIP}$ [this morning]$_{TMP}$.

Mr.$_{B-AGENT}$ Smith$_{I}$ sent the$_{B-OBJ}$ report$_{I}$ to$_{O}$ me$_{B-RECIP}$ this$_{B-TMP}$ morning$_{I}$.$_{O}$

Semantic Role Labeling: The Problem

Structural view

Mapping from input to output structures:

- **Input** is *text* (enriched with morpho-syntactic information)
- **Output** is a *sequence of labeled arguments*
- **Sequential** segmenting/labeling problem

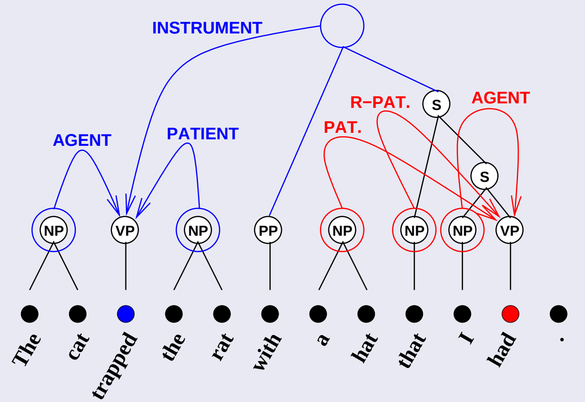
" Mr. Smith sent the report to me this morning . "

[Mr. Smith]$_{AGENT}$ sent [the report]$_{OBJ}$ to [me]$_{RECIP}$ [this morning]$_{TMP}$.

Mr.$_{B-AGENT}$ Smith$_{I}$ sent the$_{B-OBJ}$ report$_{I}$ to$_{O}$ me$_{B-RECIP}$ this$_{B-TMP}$ morning$_{I}$.$_{O}$

Semantic Role Labeling: The Problem

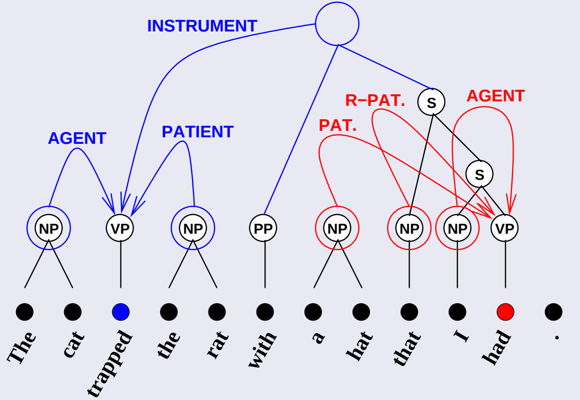
Structural View



Output is a *hierarchy of labeled arguments*

Semantic Role Labeling: The Problem

Structural View

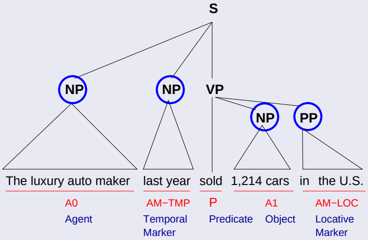


Output is a *hierarchy of labeled arguments*

Semantic Role Labeling: The Problem

Linguistic nature of the problem

- Argument identification is strongly related to syntax

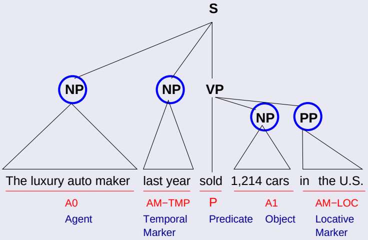


- Role labeling is a semantic task
  - e.g., selectional preferences should play an important role

Semantic Role Labeling: The Problem

Linguistic nature of the problem

- Argument identification is strongly related to syntax



- Role labeling is a semantic task
  - e.g., selectional preferences should play an important role

Semantic Role Labeling: Applications

Is SRL really useful for NLP applications?

1. Information Extraction (Surdeanu et al., 2003; Frank et al., 2007)
2. Question & Answering (Narayanan and Harabagiu, 2004)
3. Automatic Summarization (Melli et al., 2005)
4. Coreference Resolution (Ponzetto and Strube, 2006)
5. Machine Translation (Boas, 2002; Giménez and Màrquez, 2007; Wu and Fung, 2009a;2009b)
6. etc. [more on SRL and applications in the last section]

Semantic Role Labeling: Applications

Is SRL really useful for NLP applications?

1. Information Extraction (Surdeanu et al., 2003; Frank et al., 2007)
2. Question & Answering (Narayanan and Harabagiu, 2004)
3. Automatic Summarization (Melli et al., 2005)
4. Coreference Resolution (Ponzetto and Strube, 2006)
5. Machine Translation (Boas, 2002; Giménez and Màrquez, 2007; Wu and Fung, 2009a;2009b)
6. etc. [more on SRL and applications in the last section]

Semantic Role Labeling: Applications

Is SRL really useful for NLP applications?

1. Information Extraction (Surdeanu et al., 2003; Frank et al., 2007)
2. Question & Answering (Narayanan and Harabagiu, 2004)
3. Automatic Summarization (Melli et al., 2005)
4. Coreference Resolution (Ponzetto and Strube, 2006)
5. Machine Translation (Boas, 2002; Giménez and Màrquez, 2007; Wu and Fung, 2009a;2009b)
6. etc. [more on SRL and applications in the last section]

Semantic Role Labeling: Applications

Is SRL really useful for NLP applications?

1. Information Extraction (Surdeanu et al., 2003; Frank et al., 2007)
2. Question & Answering (Narayanan and Harabagiu, 2004)
3. Automatic Summarization (Melli et al., 2005)
4. Coreference Resolution (Ponzetto and Strube, 2006)
5. Machine Translation (Boas, 2002; Giménez and Màrquez, 2007; Wu and Fung, 2009a;2009b)
6. etc. [more on SRL and applications in the last section]

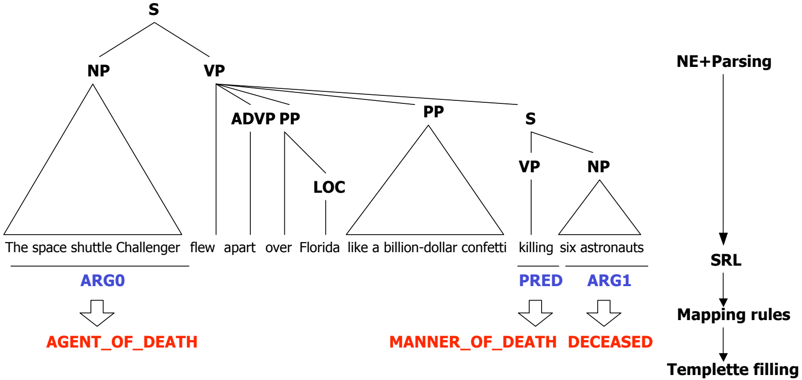
Semantic Role Labeling: Applications

Is SRL really useful for NLP applications?

1. Information Extraction (Surdeanu et al., 2003; Frank et al., 2007)
2. Question & Answering (Narayanan and Harabagiu, 2004)
3. Automatic Summarization (Melli et al., 2005)
4. Coreference Resolution (Ponzetto and Strube, 2006)
5. Machine Translation (Boas, 2002; Giménez and Màrquez, 2007; Wu and Fung, 2009a;2009b)
6. etc. [more on SRL and applications in the last section]

Walk-Through Example

The space shuttle Challenger flew apart over Florida like a billion-dollar confetti killing six astronauts.



Semantic Role Labeling: in Context

Is SRL a new problem/task?

- SRL = *shallow semantic analysis* (semantic parsing)
- Computational Semantics **is not** a **new** area in CL (actually, it is as old as AI itself)
- For decades: manual development of lexicons, grammars and other semantic resources (Hirst, 1987; Pustejovsky, 1995; Copestake & Flickinger, 2000)
- Last six years: availability of semantically annotated corpora (e.g., PropBank, FrameNet)
- Proliferation of automatic SRL systems based on statistical learning

Semantic Role Labeling: in Context

Is SRL a new problem/task?

- SRL = *shallow semantic analysis* (semantic parsing)
- Computational Semantics **is not** a **new** area in CL (actually, it is as old as AI itself)
- For decades: manual development of lexicons, grammars and other semantic resources (Hirst, 1987; Pustejovsky, 1995; Copestake & Flickinger, 2000)
- Last six years: availability of semantically annotated corpora (e.g., PropBank, FrameNet)
- Proliferation of automatic SRL systems based on statistical learning

Semantic Role Labeling: in Context

Is SRL a new problem/task?

- SRL = *shallow semantic analysis* (semantic parsing)
- Computational Semantics **is not** a **new** area in CL (actually, it is as old as AI itself)
- For decades: manual development of lexicons, grammars and other semantic resources (Hirst, 1987; Pustejovsky, 1995; Copestake & Flickinger, 2000)
- Last six years: availability of semantically annotated corpora (e.g., PropBank, FrameNet)
- Proliferation of automatic SRL systems based on statistical learning

Semantic Role Labeling: in Context

Is SRL a new problem/task?

- Other related tasks on predicate semantics (related with syntactic structure at sentence level):
  - Verb clustering according to argument structure properties (Merlo & Stevenson, 2001; Schulte im Walde, 2006)
  - Acquisition of subcategorization patterns and selectional preferences (Briscoe & Carroll, 1997)
  - Classification of semantic relations in noun phrases (Moldovan et al., 2004; Rosario & Hearst, 2004)
  - Semantic classification of prepositions (Litkowski et al., 2005)
  - Prediction of GLARF (Grammatical and Logical Representation Framework) dependency structures (Meyers et al., 2009)

Semantic Role Labeling: in Context

Is SRL a new problem/task?

- See (Yih & Toutanova, 2006) tutorial for a comparison of SRL to other related tasks and applications: Information Extraction, semantic parsing for speech dialogs and NL interfaces to DBs, deep semantic parsing, and prediction of function tags and case markers

Semantic Role Labeling: in Context

Focus of this tutorial

- We will concentrate on:

  development and learning of computational SRL systems

- Specific points
  - Statistical modeling and learning strategies
  - Resources and feature engineering
  - Evaluation and results
  - Current shortcomings and future challenges

Semantic Role Labeling: in Context

Focus of this tutorial

- We will concentrate on:

  development and learning of computational SRL systems

- Specific points
  - Statistical modeling and learning strategies
  - Resources and feature engineering
  - Evaluation and results
  - Current shortcomings and future challenges

Tutorial Overview

1. Introduction
   - Problem definition and properties
   - Main Computational Resources and Systems
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
5. Conclusions

SRL: Computational Resources

From theory to computational resources

- Since (Fillmore, 1968), considerable linguistic research has been devoted to the nature of semantic roles
- Two broad families exist:

1. Syntax-based approach : explaining the varied expression of verb arguments within syntactic positions : Levin (1993) verb classes ⟹ VerbNet (Kipper et al., 2000) ⟹ PropBank (Palmer et al., 2005) : Focused on verbs

2. Situation-based approach (a word activates/invokes a frame of semantic knowledge that relates linguistic semantics to encyclopedic knowledge) : Frame semantics (Fillmore, 1976) ⟹ FrameNet (Fillmore et al., 2004) : Words with other POS can invoke frames too (e.g., nouns, adjectives)

SRL: Computational Resources

From theory to computational resources

- Since (Fillmore, 1968), considerable linguistic research has been devoted to the nature of semantic roles
- Two broad families exist:
  1. Syntax-based approach : explaining the varied expression of verb arguments within syntactic positions : Levin (1993) verb classes $\Longrightarrow$ VerbNet (Kipper et al., 2000) $\Longrightarrow$ PropBank (Palmer et al., 2005) : Focused on verbs
  2. Situation-based approach (a word activates/invokes a frame of semantic knowledge that relates linguistic semantics to encyclopedic knowledge) : Frame semantics (Fillmore, 1976) $\Longrightarrow$ FrameNet (Fillmore et al., 2004) : Words with other POS can invoke frames too (e.g., nouns, adjectives)

SRL: Computational Resources

From theory to computational resources

- Since (Fillmore, 1968), considerable linguistic research has been devoted to the nature of semantic roles
- Two broad families exist:
  1. Syntax-based approach : explaining the varied expression of verb arguments within syntactic positions : Levin (1993) verb classes $\Longrightarrow$ VerbNet (Kipper et al., 2000) $\Longrightarrow$ PropBank (Palmer et al., 2005) : Focused on verbs
  2. Situation-based approach (a word activates/invokes a frame of semantic knowledge that relates linguistic semantics to encyclopedic knowledge) : Frame semantics (Fillmore, 1976) $\Longrightarrow$ FrameNet (Fillmore et al., 2004) : Words with other POS can invoke frames too (e.g., nouns, adjectives)

Semantic Role Labeling: Corpora

FrameNet (Fillmore et al., 2004)

- FrameNet Project: http://framenet.icsi.berkeley.edu
- Based on the theory of Semantic Frames (Fillmore, 1976)
- Methodology followed by lexicographers:
  - Define a situation based frame (e.g., Arrest)
  - Identify lexical items that invoke the frame (lexical units, e.g., "aprehend", "bust")
  - Define appropriate roles for the frame (frame elements, e.g., Suspect, Authorities, Offense)
  - Find example sentences in the corpus and annotate them

Semantic Role Labeling: Corpora

FrameNet (Fillmore et al., 2004)

- FrameNet Project: http://framenet.icsi.berkeley.edu
- Based on the theory of Semantic Frames (Fillmore, 1976)
- Methodology followed by lexicographers:
  - Define a situation based frame (e.g., Arrest)
  - Identify lexical items that invoke the frame (lexical units, e.g., “aprehend”, “bust”)
  - Define appropriate roles for the frame (frame elements, e.g., Suspect, Authorities, Offense)
  - Find example sentences in the corpus and annotate them

Semantic Role Labeling: Corpora

FrameNet (Fillmore et al., 2004)

Main characteristics

- Computational frame lexicon + corpus of examples annotated with semantic roles (mostly BNC)
  - ∼800 semantic frames
  - >9,000 lexical units
  - ∼150,000 annotated sentences
- Frame specific roles
- Corpus is not a representative sample of text

Semantic Role Labeling: Corpora

FrameNet (Fillmore et al., 2004)

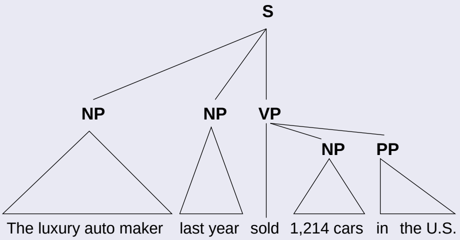
Main characteristics

- Computational frame lexicon + corpus of examples annotated with semantic roles (mostly BNC)
  - ∼800 semantic frames
  - >9,000 lexical units
  - ∼150,000 annotated sentences
- Frame specific roles
- Corpus is not a representative sample of text

Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

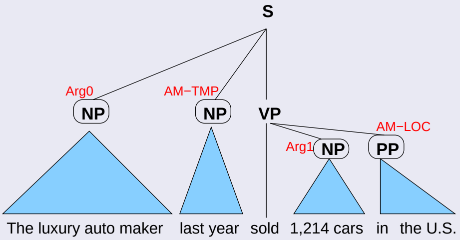
- Annotation of all verbal predicates in WSJ (Penn Treebank)
- http://verbs.colorado.edu/~mpalmer/projects/ace.html
- Add a semantic layer to the Syntactic Trees



Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

- Annotation of all verbal predicates in WSJ (Penn Treebank)
- http://verbs.colorado.edu/~mpalmer/projects/ace.html
- Add a semantic layer to the Syntactic Trees



Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

- Theory neutral numeric core roles (Arg0, Arg1, etc.)
  - Interpretation of roles: verb-specific framesets
  - Arg0 and Arg1 usually correspond to prototypical Agent and Patient/Theme roles. Other arguments do not consistently generalize across verbs
  - Different senses have different framesets
  - Syntactic alternations that preserve meaning are kept toghether in a single frameset
- Closed set of 13 general labels for Adjuncts (e.g., Temporal, Manner, Location, etc.)

Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

- Theory neutral numeric core roles (Arg0, Arg1, etc.)
  - Interpretation of roles: verb-specific framesets
  - Arg0 and Arg1 usually correspond to prototypical Agent and Patient/Theme roles. Other arguments do not consistently generalize across verbs
  - Different senses have different framesets
  - Syntactic alternations that preserve meaning are kept toghether in a single frameset
- Closed set of 13 general labels for Adjuncts (e.g., Temporal, Manner, Location, etc.)

Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

- Theory neutral numeric core roles (Arg0, Arg1, etc.)
  - Interpretation of roles: verb-specific framesets
  - Arg0 and Arg1 usually correspond to prototypical Agent and Patient/Theme roles. Other arguments do not consistently generalize across verbs
  - Different senses have different framesets
  - Syntactic alternations that preserve meaning are kept toghether in a single frameset
- Closed set of 13 general labels for Adjuncts (e.g., Temporal, Manner, Location, etc.)

Semantic Role Labeling: Corpora

PropBank: Frame files (Palmer et al., 2005)

- sell.01: commerce: seller
  Arg0="seller" (*agent*); Arg1="thing sold" (*theme*); Arg2="buyer" (*recipient*); Arg3="price paid"; Arg4="benefactive"
  [Al Brownstein]$_{\mathit{Arg0}}$ sold [it]$_{\mathit{Arg1}}$ [for $60 a bottle]$_{\mathit{Arg3}}$
- sell.02: give up
  Arg0="entity selling out"
  [John]$_{\mathit{Arg0}}$ sold out
- sell.03: sell until none is/are left
  Arg0="seller"; Arg1="thing sold"; ...
  [The new Harry Potter]$_{\mathit{Arg1}}$ sold out [within 20 minutes]$_{\mathit{ArgM-TMP}}$

Semantic Role Labeling: Corpora

PropBank: Frame files (Palmer et al., 2005)

- sell.01: commerce: seller
  Arg0="seller" (*agent*); Arg1="thing sold" (*theme*); Arg2="buyer" (*recipient*); Arg3="price paid"; Arg4="benefactive"

  [Al Brownstein]$_{Arg0}$ sold [it]$_{Arg1}$ [for $60 a bottle]$_{Arg3}$

- sell.02: give up
  Arg0="entity selling out"

  [John]$_{Arg0}$ sold out

- sell.03: sell until none is/are left
  Arg0="seller"; Arg1="thing sold"; ...

  [The new Harry Potter]$_{Arg1}$ sold out [within 20 minutes]$_{ArgM-TMP}$

Semantic Role Labeling: Corpora

PropBank: Frame files (Palmer et al., 2005)

- sell.01: commerce: seller
  Arg0="seller" (*agent*); Arg1="thing sold" (*theme*); Arg2="buyer" (*recipient*); Arg3="price paid"; Arg4="benefactive"

  [Al Brownstein]$_{Arg0}$ sold [it]$_{Arg1}$ [for \$60 a bottle]$_{Arg3}$

- sell.02: give up
  Arg0="entity selling out"

  [John]$_{Arg0}$ sold out

- sell.03: sell until none is/are left
  Arg0="seller"; Arg1="thing sold"; ...

  [The new Harry Potter]$_{Arg1}$ sold out [within 20 minutes]$_{ArgM-TMP}$

Semantic Role Labeling: Corpora

PropBank: Frame files (Palmer et al., 2005)

- sell.01: commerce: seller
  Arg0="seller" (*agent*); Arg1="thing sold" (*theme*); Arg2="buyer" (*recipient*); Arg3="price paid"; Arg4="benefactive"

  [Al Brownstein]$_{Arg0}$ sold [it]$_{Arg1}$ [for $60 a bottle]$_{Arg3}$

- sell.02: give up
  Arg0="entity selling out"

  [John]$_{Arg0}$ sold out

- sell.03: sell until none is/are left
  Arg0="seller"; Arg1="thing sold"; ...

  [The new Harry Potter]$_{Arg1}$ sold out [within 20 minutes]$_{ArgM-TMP}$

Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

Main characteristics

- Representative sample of text
  [but: limited genre of WSJ text]
- Non situation specific labels
  [but: core labels do not (completely) generalize across verbs]
- Has become the primary resource for research in SRL

Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

Main characteristics

- Representative sample of text
  [but: limited genre of WSJ text]
- Non situation specific labels
  [but: core labels do not (completely) generalize across verbs]
- Has become the primary resource for research in SRL

Semantic Role Labeling: Corpora

PropBank (Palmer et al., 2005)

Main characteristics

- Representative sample of text
  [but: limited genre of WSJ text]
- Non situation specific labels
  [but: core labels do not (completely) generalize across verbs]
- Has become the primary resource for research in SRL

Semantic Role Labeling: Corpora

NomBank (Meyers et al., 2004)

- NomBank Project: http://nlp.cs.nyu.edu/meyers/NomBank.html
- Annotation of the nominal predicates in WSJ–PennTreeBank

  IBM appointed John
  John was appointed by IBM
  IBM's appointment of John
  The appointment of John by IBM
  John is the current IBM appointee

- Annotation similar to PropBank

  [Her]$_{Arg0}$ gift of [a book]$_{Arg1}$ [to John]$_{Arg2}$

Semantic Role Labeling: Corpora

NomBank (Meyers et al., 2004)

- NomBank Project: http://nlp.cs.nyu.edu/meyers/NomBank.html
- Annotation of the nominal predicates in WSJ–PennTreeBank

  IBM appointed John
  John was appointed by IBM
  IBM's appointment of John
  The appointment of John by IBM
  John is the current IBM appointee

- Annotation similar to PropBank

  [Her]$_{Arg0}$ gift of [a book]$_{Arg1}$ [to John]$_{Arg2}$

Semantic Role Labeling: Corpora

Languages other than English

- Chinese PropBank
  http://verbs.colorado.edu/chinese/cpb/
- Korean PropBank
  http://www.ldc.upenn.edu/
- AnCora corpus: Spanish and Catalan
  http://http://clic.ub.edu/ancora/
- Prague Dependency Treebank: Czech
  http://ufal.mff.cuni.cz/pdt2.0/
- Penn Arabic TreeBank: Arabic
  http://www.ircs.upenn.edu/arabic/
- Others are under development, e.g., Scandinavian and Baltic languages

Semantic Role Labeling: Corpora

Other extensions

- FrameNet for German (SALSA corpus), Spanish and Japanese
- OntoNotes corpus: TreeBank + PropBank + word senses + coreference annotation
  http://www.bbn.com/NLP/OntoNotes
- CoNLL–2008 shared task: joint representation for syntactic and semantic dependencies
  http://www.yr-bcn.es/conll2008/
- CoNLL–2009 shared task: extension to multiple languages (Catalan, Chinese, Czech, English, German, Japanese, Spanish)
  http://ufal.mff.cuni.cz/conll2009-st/

Semantic Role Labeling: Systems Available

Tools available online that produce SRL structures

- ASSERT (Automatic Statistical SEmantic Role Tagger)
  http://cemantix.org/assert
- UIUC system
  `http://l2r.cs.uiuc.edu/~cogcomp/srl-demo.php`
- SwiRL
  http://www.surdeanu.name/mihai
- Shalmaneser: FrameNet-based system from SALSA project
  http://www.coli.uni-saarland.de/projects/salsa/shal/

Tutorial Overview

1 Introduction

2 State-of-the-art
- Architecture
- Feature engineering
- SRL systems in detail

3 Empirical evaluation and lessons learned

4 Problems and challenges

5 Conclusions

Tutorial Overview

1. Introduction
2. State-of-the-art
   - Architecture
   - Feature engineering
   - SRL systems in detail
3. Empirical evaluation and lessons learned
4. Problems and challenges
5. Conclusions

SRL: Step by Step

The Problem

- Given a sentence and a designated predicate p
- Every subsequence of words (not necessarily contiguous) is a potential argument of p
- Arguments can be discontinuous:
- SRL can be formalized as a mapping from word substrings to the set of argument labels plus 'non-argument'
- This is clearly impractical. We need to filter the set of candidates...

SRL: Step by Step

The Problem

- Given a sentence and a designated predicate p
- Every subsequence of words (not necessarily contiguous) is a potential argument of p
- Arguments can be discontinuous:
  "One troubling aspect of DEC's results, analysts said, was its performance in Europe"
- SRL can be formalized as a mapping from word substrings to the set of argument labels plus 'non-argument'
- This is clearly impractical. We need to filter the set of candidates...

SRL: Step by Step

The Problem

- Given a sentence and a designated predicate p
- Every subsequence of words (not necessarily contiguous) is a potential argument of p
- Arguments can be discontinuous:
  [One troubling aspect of DEC's results]$_{Arg1}$, [analysts]$_{Arg0}$ said, [was its performance in Europe]$_{C-Arg1}$.
- SRL can be formalized as a mapping from word substrings to the set of argument labels plus 'non-argument'
- This is clearly impractical. We need to filter the set of candidates...

SRL: Step by Step

The Problem

- Given a sentence and a designated predicate p
- Every subsequence of words (not necessarily contiguous) is a potential argument of p
- Arguments can be discontinuous:
  [One troubling aspect of DEC's results]$_{Arg1}$, [analysts]$_{Arg0}$ said, [was its performance in Europe]$_{C-Arg1}$.
- SRL can be formalized as a mapping from word substrings to the set of argument labels plus 'non-argument'
- This is clearly impractical. We need to filter the set of candidates...

SRL: Step by Step

Step 1: Select argument candidates

- Given a sentence and a designated predicate
- Parse the sentence
- Identify candidates in tree constituents (filtering/pruning)
  - Simple heuristic rules can be used, which maintain a high recall (Xue & Palmer, 2004)
- Key point: 95% of semantic arguments coincide with unique syntactic constituents in the gold parse tree (PropBank)
  - Matching is still ∼90% when using automatic parsers

SRL: Step by Step

Step 1: Select argument candidates

- Given a sentence and a designated predicate
- Parse the sentence
- Identify candidates in tree constituents (filtering/pruning)
  - Simple heuristic rules can be used, which maintain a high recall (Xue & Palmer, 2004)
- Key point: 95% of semantic arguments coincide with unique syntactic constituents in the gold parse tree (PropBank)
  - Matching is still ∼90% when using automatic parsers

SRL: Step by Step

Step 2: Local scoring of candidates

- Apply classifiers to assign confidence scores to argument candidates (all labels + 'non-argument')
- Candidates are treated independently of each other
- *Identification* and *Classification* may be performed separately
  - Computational reasons but also modularity in feature engineering
- Many ML paradigms have been used: not big differences
- Features are more important

SRL: Step by Step

Step 2: Local scoring of candidates

- Apply classifiers to assign confidence scores to argument candidates (all labels + 'non-argument')
- Candidates are treated independently of each other
- *Identification* and *Classification* may be performed separately
  - Computational reasons but also modularity in feature engineering
- Many ML paradigms have been used: not big differences
- Features are more important

SRL: Step by Step

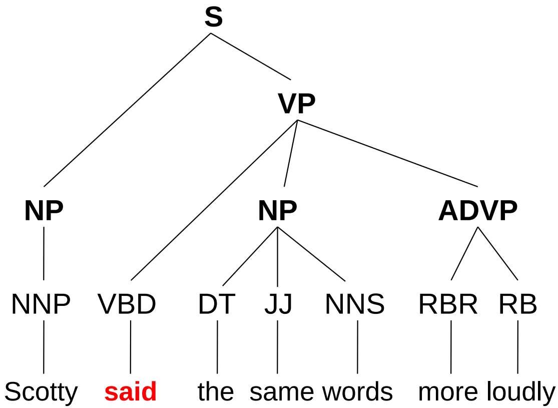
Step 2: Local scoring of candidates

- Apply classifiers to assign confidence scores to argument candidates (all labels + 'non-argument')
- Candidates are treated independently of each other
- *Identification* and *Classification* may be performed separately
  - Computational reasons but also modularity in feature engineering
- Many ML paradigms have been used: not big differences
- Features are more important

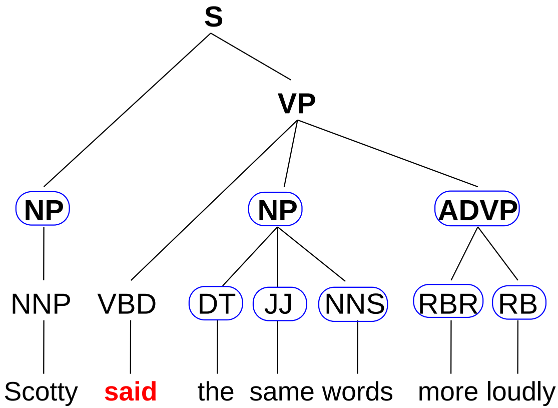
SRL: Steps 1 + 2

Scotty **said** the same words more loudly

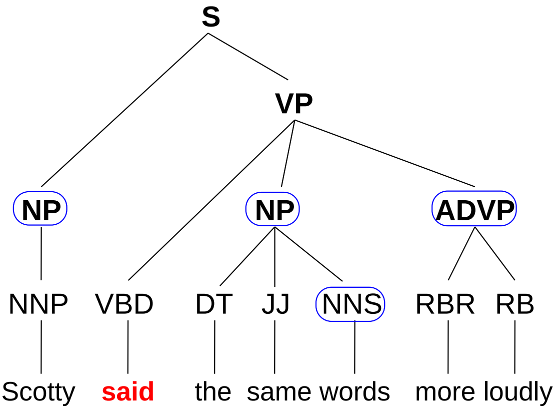
SRL: Steps 1 + 2


S
VP
NP
NP
ADVP
NNP
VBD
DT
JJ
NNS
RBR
RB
Scotty
said
the
same
words
more
loudly

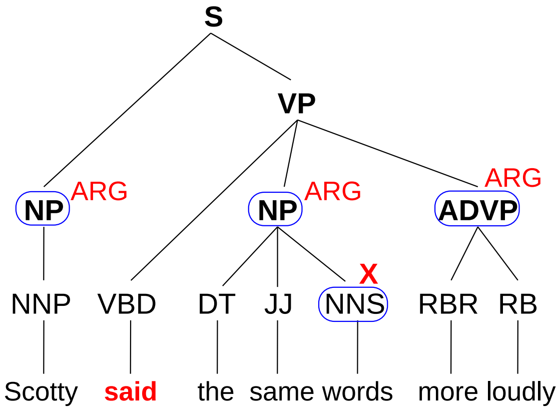
SRL: Steps 1 + 2


S
VP
NP
NP
ADVP
NNP
VBD
DT
JJ
NNS
RBR
RB
Scotty
said
the
same
words
more
loudly

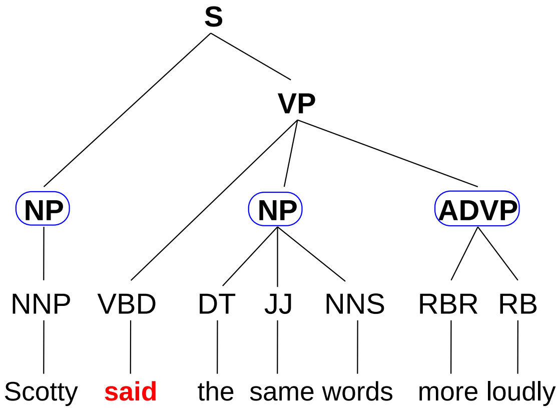
SRL: Steps 1 + 2


S
VP
NP
NP
ADVP
NNP
VBD
DT
JJ
NNS
RBR
RB
Scotty
said
the
same
words
more
loudly

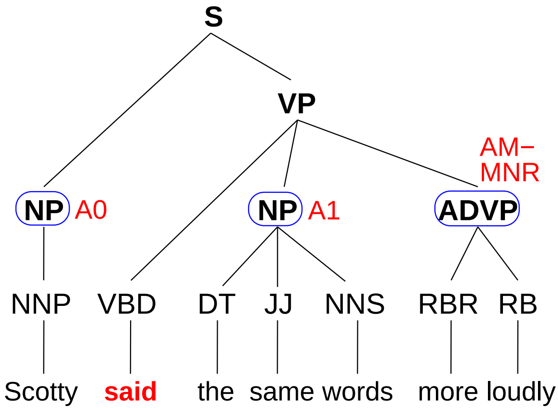
SRL: Steps 1 + 2


S
VP
NP
ARG
NP
ARG
ADVP
ARG
X
NNP
VBD
DT
JJ
NNS
RBR
RB
Scotty
said
the
same
words
more
loudly

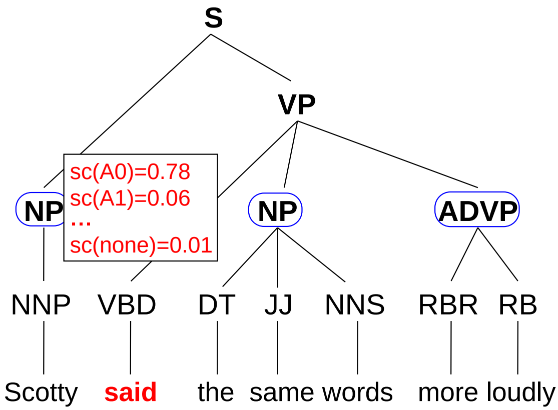
SRL: Steps 1 + 2


S
VP
NP
NP
ADVP
NNP
VBD
DT
JJ
NNS
RBR
RB
Scotty
said
the
same
words
more
loudly

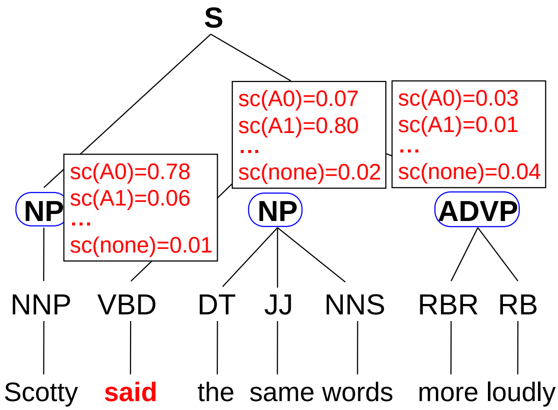
SRL: Steps 1 + 2


S
VP
AM–
MNR
NP A0
NP A1
ADVP
NNP VBD DT JJ NNS RBR RB
Scotty said the same words more loudly

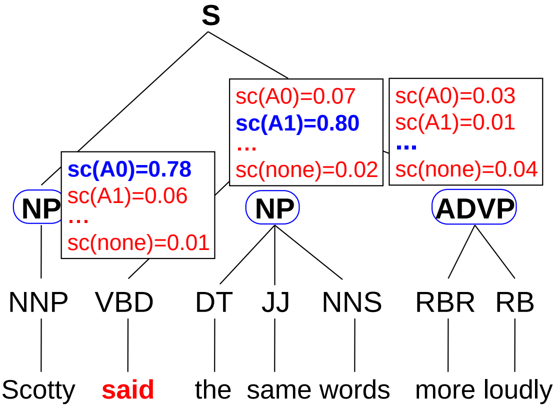
SRL: Motivating next step (joint scoring)



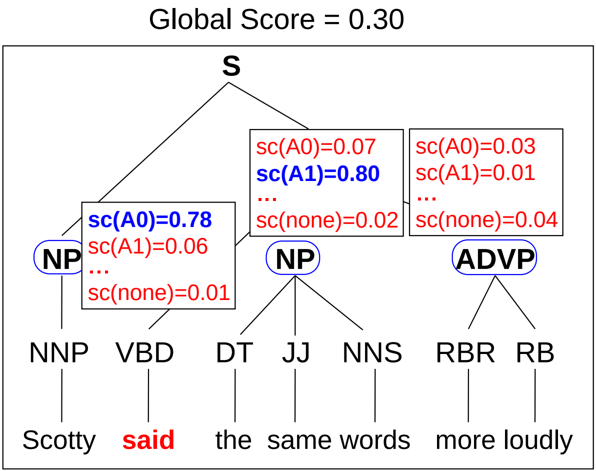
SRL: Motivating next step (joint scoring)



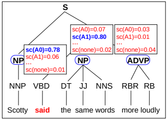
SRL: Motivating next step (joint scoring)



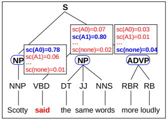
SRL: Motivating next step (joint scoring)



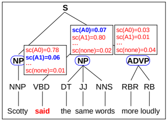
SRL: Motivating next step (joint scoring)



GS = 0.45

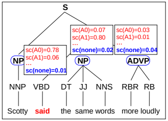


GS = 0.15



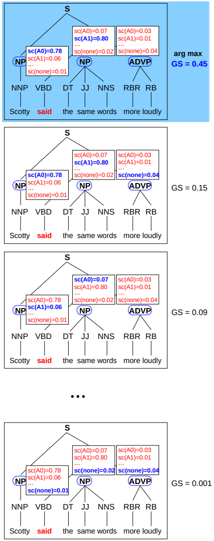
GS = 0.09

...



GS = 0.001

SRL: Motivating next step (joint scoring)



SRL: Step by Step

Step 3: Joint scoring — Paradigmatic examples

- Combine local predictions through ILP to find the best solution according to structural and linguistic constraints (Koomen et al., 2005; Punyakanok et al., 2008)

  −learning +dependencies +search

- Re-ranking of several candidate solutions (Haghighi et al., 2005; Toutanova et al., 2008)

  +learning +dependencies −search

- Global search integrating joint scoring: Tree CRFs (Cohn & Blunsom, 2005)

  +learning +/−dependencies +/−search

SRL: Step by Step

Step 3: Joint scoring — Paradigmatic examples

- Combine local predictions through ILP to find the best solution according to structural and linguistic constraints
  (Koomen et al., 2005; Punyakanok et al., 2008)

−learning +dependencies +search

- Re-ranking of several candidate solutions
  (Haghighi et al., 2005; Toutanova et al., 2008)

+learning +dependencies −search

- Global search integrating joint scoring: Tree CRFs
  (Cohn & Blunsom, 2005)

+learning +/−dependencies +/−search

SRL: Step by Step

Step 4: Post-processing

- Application of a set of heuristic rules to:

  Correct frequent errors

  Enforce consistency in the solution

SRL: Step by Step

Exceptions to the standard architecture

1. Joint treatment of all predicates in the sentence
   (Carreras et al., 2004; Surdeanu et al., 2007)
2. Specialized parsing for SRL
   - Syntactic parser trained to predict argument candidates
     (Yi & Palmer, 2005)
   - Joint parsing and SRL: semantic parsing
     (Musillo & Merlo, 2006; Merlo & Musillo, 2008)
   - SRL based on dependency parsing (Johansson & Nugues, 2007)
   - Systems from the CoNLL–2008 and 2009 shared tasks
     (Surdeanu et al., 2008; Hajič et al., 2009)
3. Sequential labeling instead of tree traversing. Motivated by:
   - The lack of full parse trees (Carreras & Màrquez, CoNLL-2004)
   - Allowing efficient search in joint inference (Màrquez et al., 2005)

SRL: Step by Step

Exceptions to the standard architecture

1. Joint treatment of all predicates in the sentence
   (Carreras et al., 2004; Surdeanu et al., 2007)
2. Specialized parsing for SRL
   - Syntactic parser trained to predict argument candidates (Yi & Palmer, 2005)
   - Joint parsing and SRL: semantic parsing (Musillo & Merlo, 2006; Merlo & Musillo, 2008)
   - SRL based on dependency parsing (Johansson & Nugues, 2007)
   - Systems from the CoNLL–2008 and 2009 shared tasks (Surdeanu et al., 2008; Hajič et al., 2009)
3. Sequential labeling instead of tree traversing. Motivated by:
   - The lack of full parse trees (Carreras & Màrquez, CoNLL–2004)
   - Allowing efficient search in joint inference (Màrquez et al., 2005)

SRL: Step by Step

Exceptions to the standard architecture

1. Joint treatment of all predicates in the sentence (Carreras et al., 2004; Surdeanu et al., 2007)
2. Specialized parsing for SRL
   - Syntactic parser trained to predict argument candidates (Yi & Palmer, 2005)
   - Joint parsing and SRL: semantic parsing (Musillo & Merlo, 2006; Merlo & Musillo, 2008)
   - SRL based on dependency parsing (Johansson & Nugues, 2007)
   - Systems from the CoNLL–2008 and 2009 shared tasks (Surdeanu et al., 2008; Hajič et al., 2009)
3. Sequential labeling instead of tree traversing. Motivated by:
   - The lack of full parse trees (Carreras & Màrquez, CoNLL-2004)
   - Allowing efficient search in joint inference (Màrquez et al., 2005)

Tutorial Overview

1. Introduction
2. State-of-the-art
   - Architecture
   - Feature engineering
   - SRL systems in detail
3. Empirical evaluation and lessons learned
4. Problems and challenges
5. Conclusions

SRL: Feature Engineering

Features: local scoring (Gildea & Jurafsky, 2002)

- Highly influential for the SRL work.
  They characterize:
  1. The candidate argument (constituent) and its context: phrase type, head word, governing category of the constituent
  2. The verb predicate and its context: lemma, voice, subcategorization pattern of the verb
  3. The relation between the consituent and the predicate: position of the constituent with respect to the verb, category path between them.

SRL: Feature Engineering

Features: local scoring (Gildea & Jurafsky, 2002)

- Highly influential for the SRL work.
  They characterize:
  1. The candidate argument (constituent) and its context: phrase type, head word, governing category of the constituent
  2. The verb predicate and its context: lemma, voice, subcategorization pattern of the verb
  3. The relation between the consituent and the predicate: position of the constituent with respect to the verb, category path between them.

SRL: Feature Engineering

Features: local scoring (Gildea & Jurafsky, 2002)

- Highly influential for the SRL work.
  They characterize:
  1. The candidate argument (constituent) and its context: phrase type, head word, governing category of the constituent
  2. The verb predicate and its context: lemma, voice, subcategorization pattern of the verb
  3. The relation between the consituent and the predicate: position of the constituent with respect to the verb, category path between them.

SRL: Feature Engineering

Features: local scoring (Gildea & Jurafsky, 2002)

- Highly influential for the SRL work.
  They characterize:
  1. The candidate argument (constituent) and its context: phrase type, head word, governing category of the constituent
  2. The verb predicate and its context: lemma, voice, subcategorization pattern of the verb
  3. The relation between the consitituent and the predicate: position of the constituent with respect to the verb, category path between them.

SRL: Feature Engineering

Features: local scoring — extensions

- "Brute force" features. Applied to the constituent and possibly to parent and siblings:
  - First and last words/POS in the constituent, bag-of-words, *n*-grams of POS, and sequence of top syntactic elements in the constituent.
- Linguistically–inspired features
  - Content word, named entities (Surdeanu et al., 2003), syntactic frame (Xue & Palmer, 2004), path variations, semantic compatibility between constituent head and predicate (Zapirain et al., 2007;2009), etc.
- Significant (and cumulative) increase in performance

SRL: Feature Engineering

Features: local scoring — extensions

- "Brute force" features. Applied to the constituent and possibly to parent and siblings:
  - First and last words/POS in the constituent, bag-of-words, *n*-grams of POS, and sequence of top syntactic elements in the constituent.
- Linguistically–inspired features
  - Content word, named entities (Surdeanu et al., 2003), syntactic frame (Xue & Palmer, 2004), path variations, semantic compatibility between constituent head and predicate (Zapirain et al., 2007;2009), etc.
- Significant (and cumulative) increase in performance

SRL: Feature Engineering

Features: local scoring — extensions

- "Brute force" features. Applied to the constituent and possibly to parent and siblings:
  - First and last words/POS in the constituent, bag-of-words, *n*-grams of POS, and sequence of top syntactic elements in the constituent.
- Linguistically–inspired features
  - Content word, named entities (Surdeanu et al., 2003), syntactic frame (Xue & Palmer, 2004), path variations, semantic compatibility between constituent head and predicate (Zapirain et al., 2007;2009), etc.
- Significant (and cumulative) increase in performance

SRL: Feature Engineering

Features: joint scoring

- Richer features taking into account information from several arguments at a time
- Best example: when doing re-ranking one may codify patterns on the whole candidate argument structure (Hiaghighi et al., 2005; Toutanova et al., 2008)
- Good for capturing global preferences
- (more on this approach in a while)

SRL: Feature Engineering

Features: joint scoring

- Richer features taking into account information from several arguments at a time
- Best example: when doing re-ranking one may codify patterns on the whole candidate argument structure (Hiaghighi et al., 2005; Toutanova et al., 2008)
- Good for capturing global preferences
- (more on this approach in a while)

SRL: Feature Engineering

Features: the Kernel approach

- Knowledge poor approach
- Let the kernel function to compute the similarity/differences between examples by considering all possible substructures as features
- Motivation: avoid intense knowledge engineering
- Potentially useful for rapid system development and working with under resourced languages
- Mostly variants of Collins' all-subtrees convolution kernel (Collins & Duffy 2001; Moschitti et al., 2008)

SRL: Feature Engineering

Features: the Kernel approach

- Knowledge poor approach
- Let the kernel function to compute the similarity/differences between examples by considering all possible substructures as features
- Motivation: avoid intense knowledge engineering
- Potentially useful for rapid system development and working with under resourced languages
- Mostly variants of Collins' all-subtrees convolution kernel (Collins & Duffy 2001; Moschitti et al., 2008)

SRL: Feature Engineering

Features: the Kernel approach

- Knowledge poor approach
- Let the kernel function to compute the similarity/differences between examples by considering all possible substructures as features
- Motivation: avoid intense knowledge engineering
- Potentially useful for rapid system development and working with under resourced languages
- Mostly variants of Collins' all-subtrees convolution kernel (Collins & Duffy 2001; Moschitti et al., 2008)

SRL: Feature Engineering

Features: the Kernel approach

Problems with the structural kernel approach

1. Uncontrolled explosion of features
2. Low efficiency
3. Inability to use linguistic knowledge

Some works in the previous directions

- Semantic Role Labeling Using a Grammar-Driven Convolution Tree Kernel. Includes approximate matching at substructure and node levels (Zhang et al., 2008)
- Feature selection in kernel space and lineariziation of Tree Kernel functions (Pighin & Moschitti, 2009)

SRL: Feature Engineering

Features: the Kernel approach

Problems with the structural kernel approach

1. Uncontrolled explosion of features
2. Low efficiency
3. Inability to use linguistic knowledge

Some works in the previous directions

- Semantic Role Labeling Using a Grammar-Driven Convolution Tree Kernel. Includes approximate matching at substructure and node levels (Zhang et al., 2008)
- Feature selection in kernel space and lineariziation of Tree Kernel functions (Pighin & Moschitti, 2009)

Tutorial Overview

1. Introduction
2. State-of-the-art
   - Architecture
   - Feature engineering
   - SRL systems in detail
3. Empirical evaluation and lessons learned
4. Problems and challenges
5. Conclusions

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Architecture

1. Identify argument candidates
   - Pruning (Xue & Palmer, 2004)
   - Argument identification: binary classification (using SNoW)
2. Classify argument candidates
   - Argument Classifier: multi-class classification (SNoW)
3. Inference
   - Use the estimated probability distribution given by the argument classifier
   - Use structural and linguistic constraints
   - Infer the optimal global output

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Architecture

1. Identify argument candidates
   - Pruning (Xue & Palmer, 2004)
   - Argument identification: binary classification (using SNoW)
2. Classify argument candidates
   - Argument Classifier: multi-class classification (SNoW)
3. Inference
   - Use the estimated probability distribution given by the argument classifier
   - Use structural and linguistic constraints
   - Infer the optimal global output

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Architecture

1. Identify argument candidates
   - Pruning (Xue & Palmer, 2004)
   - Argument identification: binary classification (using SNoW)
2. Classify argument candidates
   - Argument Classifier: multi-class classification (SNoW)
3. Inference
   - Use the estimated probability distribution given by the argument classifier
   - Use structural and linguistic constraints
   - Infer the optimal global output

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Inference

- The output of the argument classifier often violates some constraints, especially when the sentence is long
- Finding the best legitimate output is formalized as an optimization problem and solved via Integer Linear Programming (Roth & Yih, 2004)
- Input formed by:
  - The probability estimation (by the argument classifier)
  - Structural and linguistic constraints
- Allows incorporating expressive constraints (non-sequential) on the variables (the arguments types)

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Inference

- The output of the argument classifier often violates some constraints, especially when the sentence is long
- Finding the best legitimate output is formalized as an optimization problem and solved via Integer Linear Programming (Roth & Yih, 2004)
- Input formed by:
  - The probability estimation (by the argument classifier)
  - Structural and linguistic constraints
- Allows incorporating expressive constraints (non-sequential) on the variables (the arguments types)

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Inference

- The output of the argument classifier often violates some constraints, especially when the sentence is long
- Finding the best legitimate output is formalized as an optimization problem and solved via Integer Linear Programming (Roth & Yih, 2004)
- Input formed by:
  - The probability estimation (by the argument classifier)
  - Structural and linguistic constraints
- Allows incorporating expressive constraints (non-sequential) on the variables (the arguments types)

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Inference

- The output of the argument classifier often violates some constraints, especially when the sentence is long
- Finding the best legitimate output is formalized as an optimization problem and solved via Integer Linear Programming (Roth & Yih, 2004)
- Input formed by:
  - The probability estimation (by the argument classifier)
  - Structural and linguistic constraints
- Allows incorporating expressive constraints (non-sequential) on the variables (the arguments types)

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Integer Linear Programming Inference

- For each candidate argument a_i $(1 \leq i \leq n)$,
  Set up a Boolean variable: $a_{i,t}$ indicating whether a_i is classified as argument type t
- Goal is to maximize: $\sum_i \mathrm{score}(a_i = t) \cdot a_{i,t}$
  Subject to the (linear) constraints
- If $\mathrm{score}(a_i = t) = P(a_i = t)$, the objective is to find the assignment that maximizes the expected number of arguments that are correct and satisfies the constraints

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Constraints: examples

- No duplicate argument classes: $\sum_{i=1}^{n} a_{i,Arg0} \leq 1$
- On discontinuous arguments (C-ARG)
  $\forall j(1 \leq j \leq n), \sum_{i=1}^{j-1} a_{i,Arg0} \geq a_{j,C-Arg0}$
- On reference arguments (R-ARG)
  $\forall j(1 \leq j \leq n), \sum_{i \neq j} a_{i,Arg0} \geq a_{j,R-Arg0}$
- Many other possible constraints:
  - Unique labels
  - No overlapping or embedding
  - Relations between number of arguments; order constraints
  - If verb is of type A, no argument of type B
- ILP inference can be used to combine different SRL systems

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Constraints: examples

- No duplicate argument classes: $\sum_{i=1}^{n} a_{i,Arg0} \leq 1$
- On discontinuous arguments (C-ARG)
  $\forall j(1 \leq j \leq n), \sum_{i=1}^{j-1} a_{i,Arg0} \geq a_{j,C-Arg0}$
- On reference arguments (R-ARG)
  $\forall j(1 \leq j \leq n), \sum_{i \neq j} a_{i,Arg0} \geq a_{j,R-Arg0}$
- Many other possible constraints:
  - Unique labels
  - No overlapping or embedding
  - Relations between number of arguments; order constraints
  - If verb is of type A, no argument of type B
- ILP inference can be used to combine different SRL systems

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Constraints: examples

- No duplicate argument classes: $\sum_{i=1}^{n} a_{i,Arg0} \leq 1$
- On discontinuous arguments (C-ARG)
  $\forall j (1 \leq j \leq n), \sum_{i=1}^{j-1} a_{i,Arg0} \geq a_{j,C-Arg0}$
- On reference arguments (R-ARG)
  $\forall j (1 \leq j \leq n), \sum_{i \neq j} a_{i,Arg0} \geq a_{j,R-Arg0}$
- Many other possible constraints:
  - Unique labels
  - No overlapping or embedding
  - Relations between number of arguments; order constraints
  - If verb is of type A, no argument of type B
- ILP inference can be used to combine different SRL systems

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Constraints: examples

- No duplicate argument classes: $\sum_{i=1}^{n} a_{i,Arg0} \leq 1$
- On discontinuous arguments (C-ARG)
  $\forall j(1 \leq j \leq n),\ \sum_{i=1}^{j-1} a_{i,Arg0} \geq a_{j,C-Arg0}$
- On reference arguments (R-ARG)

  [The deregulation]$_{Arg1}$ of railroads and trucking companies [that]$_{R-Arg1}$ began [in 1980]$_{AM-TMP}$ enabled ...

  $\forall j(1 \leq j \leq n),\ \sum_{i \neq j} a_{i,Arg0} \geq a_{j,R-Arg0}$
- Many other possible constraints:
  - Unique labels
  - No overlapping or embedding
  - Relations between number of arguments; order constraints
  - If verb is of type A, no argument of type B
- ILP inference can be used to combine different SRL systems

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Constraints: examples

- No duplicate argument classes: $\sum_{i=1}^{n} a_{i,Arg0} \leq 1$
- On discontinuous arguments (C-ARG)
  $\forall j (1 \leq j \leq n),\ \sum_{i=1}^{j-1} a_{i,Arg0} \geq a_{j,C-Arg0}$
- On reference arguments (R-ARG)
  $\forall j (1 \leq j \leq n),\ \sum_{i \neq j} a_{i,Arg0} \geq a_{j,R-Arg0}$
- Many other possible constraints:
  - Unique labels
  - No overlapping or embedding
  - Relations between number of arguments; order constraints
  - If verb is of type A, no argument of type B
- ILP inference can be used to combine different SRL systems

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Constraints: examples

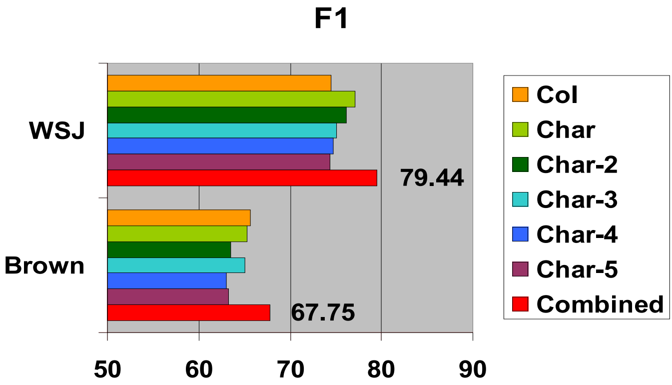
- No duplicate argument classes: $\sum_{i=1}^{n} a_{i,Arg0} \leq 1$
- On discontinuous arguments (C-ARG)
  $\forall j(1 \leq j \leq n), \sum_{i=1}^{j-1} a_{i,Arg0} \geq a_{j,C-Arg0}$
- On reference arguments (R-ARG)
  $\forall j(1 \leq j \leq n), \sum_{i \neq j} a_{i,Arg0} \geq a_{j,R-Arg0}$
- Many other possible constraints:
  - Unique labels
  - No overlapping or embedding
  - Relations between number of arguments; order constraints
  - If verb is of type A, no argument of type B
- ILP inference can be used to combine different SRL systems

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)

Constraints: examples

- No duplicate argument classes: $\sum_{i=1}^{n} a_{i,Arg0} \leq 1$
- On discontinuous arguments (C-ARG)
  $\forall j (1 \leq j \leq n),\ \sum_{i=1}^{j-1} a_{i,Arg0} \geq a_{j,C-Arg0}$
- On reference arguments (R-ARG)
  $\forall j (1 \leq j \leq n),\ \sum_{i \neq j} a_{i,Arg0} \geq a_{j,R-Arg0}$
- Many other possible constraints:
  - Unique labels
  - No overlapping or embedding
  - Relations between number of arguments; order constraints
  - If verb is of type A, no argument of type B
- ILP inference can be used to combine different SRL systems

Generalized Inference – ILP (Koomen et al., 2005; Punyakanok et al., 2008)



- Inference with many parsers improves results ~ 2.6 F_1 points
- Best results at CoNLL-2005 shared task (Carreras & Màrquez, 2005)

Joint System based on Reranking (Toutanova et al., 2008)

Architecture

- Use a probabilistic local SRL model to produce multiple (n-best) candidate solutions for the predicate structure
- Use a feature–rich reranking model to select the best solution among them
- Main goal: is to build a rich model for joint scoring, which takes into account the dependencies among the labels of argument phrases
- Use a second layer of reranking by combining different solutions coming from alternatieve input syntactic parses

Joint System based on Reranking (Toutanova et al., 2008)

Architecture

- Use a probabilistic local SRL model to produce multiple (*n*-best) candidate solutions for the predicate structure
- Use a feature–rich reranking model to select the best solution among them
- Main goal: is to build a rich model for joint scoring, which takes into account the dependencies among the labels of argument phrases
- Use a second layer of reranking by combining different solutions coming from alternatieve input syntactic parses

Joint System based on Reranking (Toutanova et al., 2008)

Models

- Simple local scoring model with strong independence assumptions, trained with log-linear models (MaxEnt): $P(labels|tree) = \prod_{node_i \in tree} P(labels_i|node_i)$
- Find top n non-overlapping assignments for local model using dynamic programming
- Select the best assignment among top n using a joint log-linear model (Collins, 2000)
- The resulting probability of a complete labeling L of the tree for a predicate p is given by: $P_{SRL}(L|tree, p) = log(P_{JOINT}(L|tree, p)) + \lambda log(P_{LOCAL}(L|tree, p))$

Joint System based on Reranking (Toutanova et al., 2008)

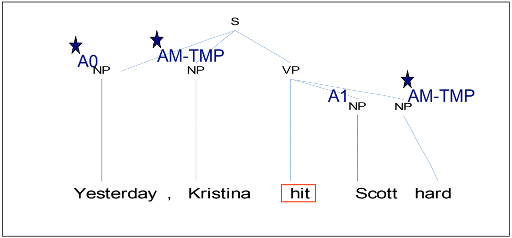
Models

- Simple local scoring model with strong independence assumptions, trained with log-linear models (MaxEnt): $P(labels|tree) = \prod_{node_i \in tree} P(labels_i|node_i)$
- Find top n non-overlapping assignments for local model using dynamic programming
- Select the best assignment among top n using a joint log-linear model (Collins, 2000)
- The resulting probability of a complete labeling L of the tree for a predicate p is given by: $P_{SRL}(L|tree, p) = log(P_{JOINT}(L|tree, p)) + \lambda log(P_{LOCAL}(L|tree, p))$

Joint System based on Reranking (Toutanova et al., 2008)

Features: joint scoring slide from (Yih & Toutanova, 2006)

Joint Model Features



Repetition features: count of arguments with a given label c(AM-TMP)=2

Complete sequence syntactic-semantic features for the core arguments:

[NP_A0 hit NP_A1] , [NP_A0 VBD NP_A1] (backoff)

[NP_A0 hit] (left backoff)

[NP_ARG hit NP_ARG] (no specific labels)

[1 hit 1] (counts of left and right core arguments)

Joint System based on Reranking (Toutanova et al., 2008)

Enhancement by using multiple trees

- For top k trees from Charniak's parser, $t_1, t_2, \ldots, t_k$, find corresponding best SRL assignments $L_1, L_2, \ldots, L_k$ and choose the tree and assignment that maximize the score (approx. joint probability of tree and assignment)
  $score(L_i, t_i) = \alpha log(P(t_i)) + log(P_{SRL}(L_i|t_i))$
- Final Results (2nd best at CoNLL):
  WSJ-23: 78.45 (F1), 79.54 (Prec.), 77.39 (Rec.)
  Brown: 67.71 (F1), 70.24 (Prec.), 65.37 (Rec.)
- Improvement due to the joint model: >2 F_1 points

Joint System based on Reranking (Toutanova et al., 2008)

Enhancement by using multiple trees

- For top k trees from Charniak's parser, $t_1, t_2, \ldots, t_k$, find corresponding best SRL assignments $L_1, L_2, \ldots, L_k$ and choose the tree and assignment that maximize the score (approx. joint probability of tree and assignment)
  $score(L_i, t_i) = \alpha log(P(t_i)) + log(P_{SRL}(L_i|t_i))$
- Final Results (2nd best at CoNLL):
  WSJ-23: 78.45 (F1), 79.54 (Prec.), 77.39 (Rec.)
  Brown: 67.71 (F1), 70.24 (Prec.), 65.37 (Rec.)
- Improvement due to the joint model: >2 F_1 points

Joint System based on Reranking (Toutanova et al., 2008)

Enhancement by using multiple trees

- For top k trees from Charniak's parser, $t_1, t_2, \ldots, t_k$, find corresponding best SRL assignments $L_1, L_2, \ldots, L_k$ and choose the tree and assignment that maximize the score (approx. joint probability of tree and assignment)
$score(L_i, t_i) = \alpha log(P(t_i)) + log(P_{SRL}(L_i|t_i))$
- Final Results (2nd best at CoNLL):
WSJ-23: 78.45 (F1), 79.54 (Prec.), 77.39 (Rec.)
Brown: 67.71 (F1), 70.24 (Prec.), 65.37 (Rec.)
- Improvement due to the joint model: >2 F_1 points

State-of-the-art: Other Systems, Approaches, etc.

- SRL using different syntactic parsers:
  - CCG parser (Gildea and Hockenmaier, 2005; Boxwell et al., 2009)
  - HPSG parsers with handcrafted grammars (Zhang et al., 2008; 2009)
- SRL using Markov Logic (Meza-Ruiz & Riedel, 2008; 2009)
- Unsupervised approaches to SRL (Swier & Stevenson, 2004;2005; Grenager & Manning, 2006; Abend et al., 2009)
- Corpora development: cross-lingual annotation projection (Fung & Chen, 2004; Padó & Lapata 2006; Fung et al., 2007; Padó 2007; Padó & Pitel, 2007)

State-of-the-art: Other Systems, Approaches, etc.

- SRL using different syntactic parsers:
  - CCG parser (Gildea and Hockenmaier, 2005; Boxwell et al., 2009)
  - HPSG parsers with handcrafted grammars (Zhang et al., 2008; 2009)
- SRL using Markov Logic (Meza-Ruiz & Riedel, 2008; 2009)
- Unsupervised approaches to SRL (Swier & Stevenson, 2004;2005; Grenager & Manning, 2006; Abend et al., 2009)
- Corpora development: cross-lingual annotation projection (Fung & Chen, 2004; Padó & Lapata 2006; Fung et al., 2007; Padó 2007; Padó & Pitel, 2007)

State-of-the-art: Other Systems, Approaches, etc.

- SRL using different syntactic parsers:
  - CCG parser (Gildea and Hockenmaier, 2005; Boxwell et al., 2009)
  - HPSG parsers with handcrafted grammars (Zhang et al., 2008; 2009)
- SRL using Markov Logic (Meza-Ruiz & Riedel, 2008; 2009)
- Unsupervised approaches to SRL (Swier & Stevenson, 2004;2005; Grenager & Manning, 2006; Abend et al., 2009)
- Corpora development: cross-lingual annotation projection (Fung & Chen, 2004; Padó & Lapata 2006; Fung et al., 2007; Padó 2007; Padó & Pitel, 2007)

State-of-the-art: Other Systems, Approaches, etc.

- SRL using different syntactic parsers:
  - CCG parser (Gildea and Hockenmaier, 2005; Boxwell et al., 2009)
  - HPSG parsers with handcrafted grammars (Zhang et al., 2008; 2009)
- SRL using Markov Logic (Meza-Ruiz & Riedel, 2008; 2009)
- Unsupervised approaches to SRL (Swier & Stevenson, 2004;2005; Grenager & Manning, 2006; Abend et al., 2009)
- Corpora development: cross-lingual annotation projection (Fung & Chen, 2004; Padó & Lapata 2006; Fung et al., 2007; Padó 2007; Padó & Pitel, 2007)

Tutorial Overview

1. Introduction
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
5. Conclusions

Empirical Evaluation of SRL Systems

Evaluation Exercises

- Up to 9 evaluation exercises in the last 5 years
  - CoNLL-2004/2005 shared tasks (Carreras & Màrquez, 2004; 2005)
  - Senseval–3 (Litkowski, 2004)
  - SemEval-2007 (Pradhan et al., 2007; Màrquez et al., 2007) (Baker et al., 2007; Litkowski & Hargraves, 2007)
  - CoNLL-2008 shared task (Surdeanu et al., 2008)
  - CoNLL-2009 shared task (Hajič et al., 2009)

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2004/2005 shared tasks

- Input: words, POS, NEs, syntax; Output: SRL annotation
- CoNLL-2004 $\Longrightarrow$ CoNLL-2005:
  - 10 teams $\Longrightarrow$ 19 teams
  - partial parsing $\Longrightarrow$ full parsing
  - $\sim$200Kw training $\Longrightarrow$ $\sim$1Mw training
- Best overall results: $\sim$80% F_1 measure
- Identifying arguments is more difficult than classifying them: recall $\sim$81%, class. accuracy $\sim$95% on the previous set
- Core arguments vs. Adjuncts: 70%–90% vs. <60%
- "Good" results on unseen predicates: $\sim$70% F_1

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2004/2005 shared tasks

- Input: words, POS, NEs, syntax; Output: SRL annotation
- CoNLL-2004 $\Longrightarrow$ CoNLL-2005:
  - 10 teams $\Longrightarrow$ 19 teams
  - partial parsing $\Longrightarrow$ full parsing
  - $\sim$200Kw training $\Longrightarrow$ $\sim$1Mw training
- Best overall results: $\sim$80% F_1 measure
- Identifying arguments is more difficult than classifying them: recall $\sim$81%, class. accuracy $\sim$95% on the previous set
- Core arguments vs. Adjuncts: 70%–90% vs. <60%
- "Good" results on unseen predicates: $\sim$70% F_1

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2004/2005 shared tasks

- Input: words, POS, NEs, syntax; Output: SRL annotation
- CoNLL-2004 $\Longrightarrow$ CoNLL-2005:
  - 10 teams $\Longrightarrow$ 19 teams
  - partial parsing $\Longrightarrow$ full parsing
  - $\sim$200Kw training $\Longrightarrow$ $\sim$1Mw training
- Best overall results: $\sim$80% F_1 measure
- Identifying arguments is more difficult than classifying them: recall $\sim$81%, class. accuracy $\sim$95% on the previous set
- Core arguments vs. Adjuncts: 70%–90% vs. <60%
- "Good" results on unseen predicates: $\sim$70% F_1

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2004/2005 shared tasks

- Input: words, POS, NEs, syntax; Output: SRL annotation
- CoNLL-2004 $\Longrightarrow$ CoNLL-2005:
  - 10 teams $\Longrightarrow$ 19 teams
  - partial parsing $\Longrightarrow$ full parsing
  - $\sim$200Kw training $\Longrightarrow$ $\sim$1Mw training
- Best overall results: $\sim$80% F_1 measure
- Identifying arguments is more difficult than classifying them: recall $\sim$81%, class. accuracy $\sim$95% on the previous set
- Core arguments vs. Adjuncts: 70%–90% vs. <60%
- "Good" results on unseen predicates: $\sim$70% F_1

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2005: System Combination

- Observation: the 4 best scoring systems at CoNLL-2005 were combined systems
- Main reason: combination increases diversity and gets more robustness from parsing errors
- What to combine? The output of different SRL base systems vs. several outputs from the same system trained using different input settings (e.g., using different parse trees)
- Combination scheme: ranking of complete solutions vs. combining argument candidates
- Combination improves results 2∼5 F_1 points

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2005: System Combination

- Observation: the 4 best scoring systems at CoNLL-2005 were combined systems
- Main reason: combination increases diversity and gets more robustness from parsing errors
- What to combine? The output of different SRL base systems vs. several outputs from the same system trained using different input settings (e.g., using different parse trees)
- Combination scheme: ranking of complete solutions vs. combining argument candidates
- $\boxed{\text{Combination improves results } 2\sim5 \ F_1 \text{ points}}$

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2005: System Combination

- Observation: the 4 best scoring systems at CoNLL-2005 were combined systems
- Main reason: combination increases diversity and gets more robustness from parsing errors
- What to combine? The output of different SRL base systems vs. several outputs from the same system trained using different input settings (e.g., using different parse trees)
- Combination scheme: ranking of complete solutions vs. combining argument candidates
- Combination improves results 2∼5 F_1 points

Empirical Evaluation: on PropBank

On PropBank: CoNLL-2005: System Combination

- Observation: the 4 best scoring systems at CoNLL-2005 were combined systems
- Main reason: combination increases diversity and gets more robustness from parsing errors
- What to combine? The output of different SRL base systems vs. several outputs from the same system trained using different input settings (e.g., using different parse trees)
- Combination scheme: ranking of complete solutions vs. combining argument candidates
- Combination improves results 2∼5 F_1 points

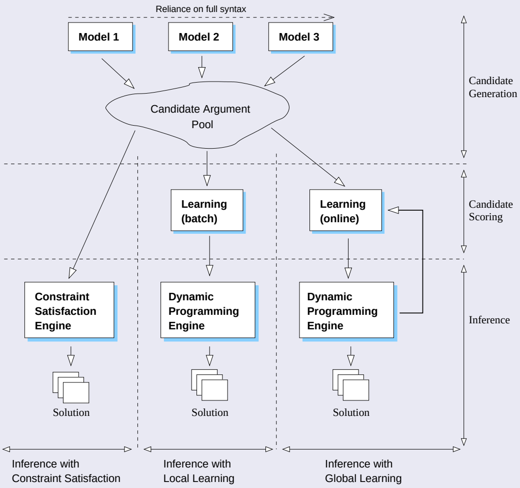
Empirical Evaluation: on PropBank

On PropBank: CoNLL-2005: System Combination

- Observation: the 4 best scoring systems at CoNLL-2005 were combined systems
- Main reason: combination increases diversity and gets more robustness from parsing errors
- What to combine? The output of different SRL base systems vs. several outputs from the same system trained using different input settings (e.g., using different parse trees)
- Combination scheme: ranking of complete solutions vs. combining argument candidates
- $\boxed{\text{Combination improves results } 2{\sim}5\ F_1 \text{ points}}$

Empirical Evaluation: on PropBank

System Combination (Surdeanu et al., 2007)



Empirical Evaluation: on PropBank

System Combination (Surdeanu et al., 2007)

Combining *n*-best systems from CoNLL-2005

	Local ranker			
WSJ	PProps	Prec.	Recall	F_1
C2	50.69%	86.60%	73.90%	79.75±0.7
C4	55.14%	86.67%	**76.63%**	81.38±0.7
C6	**54.85%**	87.45%	76.34%	**81.52**±0.6
C8	54.36%	**87.49%**	76.12%	81.41±0.6
C10	53.90%	87.48%	75.81%	81.23±0.6

Best results up to date on CoNLL-2005 datasets

Empirical Evaluation: on PropBank

On PropBank: SemEval-2007 Task #17 (Pradhan et al., 2007)

- SRL + WSD in a set of 50 selected verbal predicates
- Double annotation and evaluation: comparison of the PropBank roleset with a VerbNet-based roleset containing general semantic roles
- Only two participant systems
- Results consistent with CoNLL-2005
- Systems predicted VerbNet-based roles as accurately as PropBank roles

Empirical Evaluation: on PropBank

On PropBank: SemEval-2007 Task #17 (Pradhan et al., 2007)

- SRL + WSD in a set of 50 selected verbal predicates
- Double annotation and evaluation: comparison of the PropBank roleset with a VerbNet-based roleset containing general semantic roles
- Only two participant systems
- Results consistent with CoNLL-2005
- Systems predicted VerbNet-based roles as accurately as PropBank roles

Empirical Evaluation: on FrameNet

On FrameNet: Senseval–3 (Litkowski, 2004)

- Replicating the experimental setting of Gildea & Jurafsky (2002)
- Subset of 40 selected frames
- Simple task (Role Classification): best result ∼92%
- Complete SRL task: best result ∼83%

Empirical Evaluation: on FrameNet

On FrameNet: Senseval–3 (Litkowski, 2004)

- Replicating the experimental setting of Gildea & Jurafsky (2002)
- Subset of 40 selected frames
- Simple task (Role Classification): best result ∼92%
- Complete SRL task: best result ∼83%

Empirical Evaluation: on FrameNet

On FrameNet: SemEval-2007 Task #19 (Baker et al., 2007)

- Realistic Setting:
  - Label running text with FrameNet semantic roles
  - Output a graph representation of the sentence semantics
  - Test was newly annotated material: contained some new frames and roles not in the FrameNet lexicon
- Three teams submitted results
- Precision percentages in the 60s
  but recall percentages in the 30s

Empirical Evaluation: on FrameNet

On FrameNet: SemEval-2007 Task #19 (Baker et al., 2007)

- Realistic Setting:
  - Label running text with FrameNet semantic roles
  - Output a graph representation of the sentence semantics
  - Test was newly annotated material: contained some new frames and roles not in the FrameNet lexicon
- Three teams submitted results
- Precision percentages in the 60s
  but recall percentages in the 30s

Empirical Evaluation: other Languages

Other Languages at SemEval-2007

- Spanish and Catalan: Task #9: only 2 participants
- Arabic: Task #19: no participants in SRL
- Czech: Task #3: cancelled

Empirical Evaluation: other Languages

Other Languages at SemEval-2007

- Spanish and Catalan: Task #9: only 2 participants
- Arabic: Task #19: no participants in SRL
- Czech: Task #3: cancelled

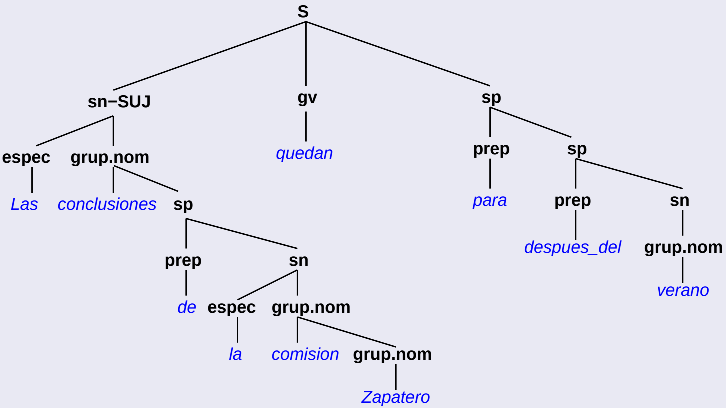
Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan

- Multilevel Semantic Annotation of Catalan and Spanish
  http://www.lsi.upc.edu/∼nlp/semeval/msacs.html
  (Màrquez et al., 2007)

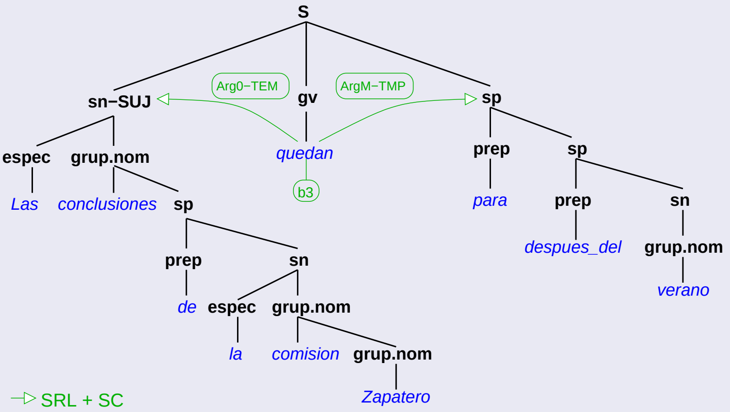
Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan



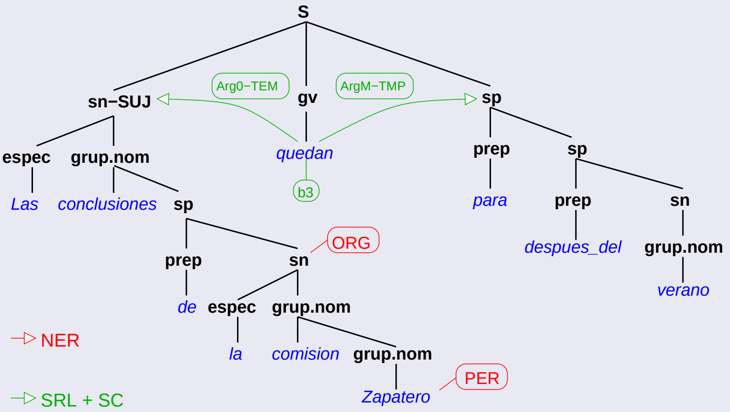
Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan



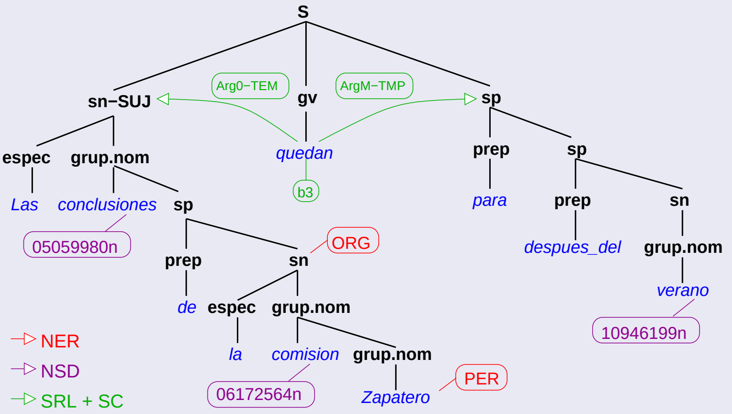
Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan



Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan



Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan

- Multilevel Semantic Annotation of Catalan and Spanish
- Goal: Joint resolution of all three semantic tasks, exploiting interdependencies among them
- Results: Best system (from ILK) showed that SRL for Catalan and Spanish is possible with comparable accuracy to state-of-the-art English systems (using gold parse trees)
- But: Nobody tried the joint learning challenge!

Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan

- Multilevel Semantic Annotation of Catalan and Spanish
- Goal: Joint resolution of all three semantic tasks, exploiting interdependencies among them
- Results: Best system (from ILK) showed that SRL for Catalan and Spanish is possible with comparable accuracy to state-of-the-art English systems (using gold parse trees)
- But: Nobody tried the joint learning challenge!

Empirical Evaluation: other Languages

SemEval-2007: Task #9 on Spanish and Catalan

- Multilevel Semantic Annotation of Catalan and Spanish
- Goal: Joint resolution of all three semantic tasks, exploiting interdependencies among them
- Results: Best system (from ILK) showed that SRL for Catalan and Spanish is possible with comparable accuracy to state-of-the-art English systems (using gold parse trees)
- But: Nobody tried the joint learning challenge!

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task

- Joint parsing of syntactic and semantic dependencies
  http://www.yr-bcn.es/conll2008/
  (Surdeanu et al., 2008)
- Main Features:
  - SRL using a dependency-based representation
  - Not only verbal predicates (from PropBank) but also nominal predicates (from NomBank)
  - More complex syntactic dependencies
  - Merged representation for syntax and semantics

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task

- Joint parsing of syntactic and semantic dependencies
  http://www.yr-bcn.es/conll2008/
  (Surdeanu et al., 2008)
- Main Features:
  - SRL using a dependency-based representation
  - Not only verbal predicates (from PropBank) but also nominal predicates (from NomBank)
  - More complex syntactic dependencies
  - Merged representation for syntax and semantics

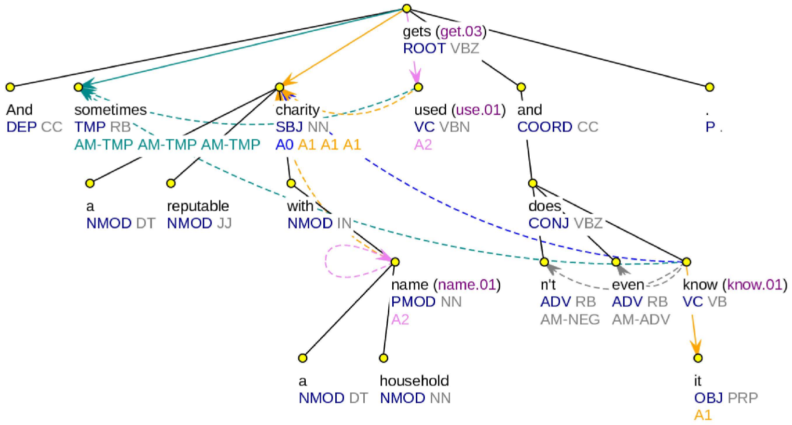
Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task

- Research questions:
  - Is the dependency-based representation better for SRL than the constituent-based formalism?
  - Is the merged representation more helpful than the individual ones?
- More motivations:
  - Ease adoption of NLP parsing technology: linear time processing possible (good fit for applications)
  - identifying the semantic dependencies between predicates and modifiers (heads of semantic arguments) could be easier and enough for application needs

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task: Graphical representation of data



Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task: some details

- Main difficulties:
  - *Input*: words + POS; *Output*: dependency tree, predicate identification and disambiguation (sense in the frame file), SRL for all predicates
  - Semantic structure does not match the syntactic dependency tree (nor any known graph representation with fast inference and learning algorithms) $\Longrightarrow$ difficult to devise joint systems
- Open/close challenges
- Full task vs. SRL-only
- Main evaluation score: global measure as a weighted average of LAS (syntax) and semantic F_1

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task: some details

- Main difficulties:
  - *Input*: words + POS; *Output*: dependency tree, predicate identification and disambiguation (sense in the frame file), SRL for all predicates
  - Semantic structure does not match the syntactic dependency tree (nor any known graph representation with fast inference and learning algorithms) $\Longrightarrow$ difficult to devise joint systems
- Open/close challenges
- Full task vs. SRL-only
- Main evaluation score: global measure as a weighted average of LAS (syntax) and semantic F_1

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task: Results and Conclusions

- 55 groups signed up for the task; 22 submitted results
- Best results (Johanson & Nugues, 2008):
  - WSJ: LAS=90.13; F_1=81.75; Overall: 85.95
  - Brown: LAS=82.81; F_1=69.06; Overall: 75.95
- Mostly pipeline architectures. 5 systems combined the syntactic and semantic subtasks to some extent (the best-performing system, among others).
  But only 2 were truly joint systems
- The best of such scored 80.19 (WSJ) and 70.34 (Brown) (Henderson et al., 2008); 5 points below the best system

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task: Results and Conclusions

- 55 groups signed up for the task; 22 submitted results
- Best results (Johanson & Nugues, 2008):
  - WSJ: LAS=90.13; F_1=81.75; Overall: 85.95
  - Brown: LAS=82.81; F_1=69.06; Overall: 75.95
- Mostly pipeline architectures. 5 systems combined the syntactic and semantic subtasks to some extent (the best-performing system, among others).
  But only 2 were truly joint systems
- The best of such scored 80.19 (WSJ) and 70.34 (Brown) (Henderson et al., 2008); 5 points below the best system

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task: On the research questions

- Comparison to CoNLL-2005:
  - Results on the dependency representation are slightly better than those on constituents. Fair post-competition comparison by Johansson (2008)
- Observation from systems addressing syntax and SRL jointly:
  - (compared to the pipeline approach) Joint inference seems not to degrade syntactic results, but to boost the F_1 score on semantic dependencies
    (Henderson et al., 2008; Lluís & Màrquez, 2008)

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2008 shared task: On the research questions

- Comparison to CoNLL-2005:
  - Results on the dependency representation are slightly better than those on constituents. Fair post-competition comparison by Johansson (2008)
- Observation from systems addressing syntax and SRL jointly:
  - (compared to the pipeline approach) Joint inference seems not to degrade syntactic results, but to boost the F_1 score on semantic dependencies (Henderson et al., 2008; Lluís & Màrquez, 2008)

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task

- Syntactic and Semantic Dependencies in Multiple Languages http://ufal.mff.cuni.cz/conll2009-st/ (Hajič et al., 2009)
- Very similar task setting and goals to those of 2008
- Particularities
  - Extension to 7 languages from different typologies: Catalan, Chinese, Czech, English, German, Japanese, Spanish
  - Significant differences among languages (e.g, corpora size, avg. sentence length, size and granularity of the syntactic and semantic tagsets, etc.)
  - Results on all languages had to be submitted
  - "Predicates" identified both in training and test

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task

- Syntactic and Semantic Dependencies in Multiple Languages http://ufal.mff.cuni.cz/conll2009-st/ (Hajič et al., 2009)
- Very similar task setting and goals to those of 2008
- Particularities
  - Extension to 7 languages from different typologies: Catalan, Chinese, Czech, English, German, Japanese, Spanish
  - Significant differences among languages (e.g, corpora size, avg. sentence length, size and granularity of the syntactic and semantic tagsets, etc.)
  - Results on all languages had to be submitted
  - “Predicates” identified both in training and test

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task: Results and Conclusions

- 68 registrations, 34 licenses for evaluation data, 20 groups submitted results
- Results:
  - Macro avg.: LAS=85.77; F_1=80.47; Overall: 82.64
  - At least one team per language beat the state-of-the-art syntactic parser provided by organizers
  - Best result on English from 2008, overall F_1=85.95 (Johansson & Nugues), was beat by 4 systems in 2009 (with best F_1=87.69)
- One surprise (about the lack of surprises): no significant changes in results among languages

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task: Results and Conclusions

- 68 registrations, 34 licenses for evaluation data, 20 groups submitted results
- Results:
  - Macro avg.: LAS=85.77; F_1=80.47; Overall: 82.64
  - At least one team per language beat the state-of-the-art syntactic parser provided by organizers
  - Best result on English from 2008, overall F_1=85.95 (Johansson & Nugues), was beat by 4 systems in 2009 (with best F_1=87.69)
- One surprise (about the lack of surprises): no significant changes in results among languages

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task: Results and Conclusions

- 68 registrations, 34 licenses for evaluation data, 20 groups submitted results
- Results:
  - Macro avg.: LAS=85.77; F_1=80.47; Overall: 82.64
  - At least one team per language beat the state-of-the-art syntactic parser provided by organizers
  - Best result on English from 2008, overall F_1=85.95 (Johansson & Nugues), was beat by 4 systems in 2009 (with best F_1=87.69)
- One surprise (about the lack of surprises): no significant changes in results among languages

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task: Results and Conclusions

- System Architectures
  - Best systems are still pipelined (syntax, then semantics)
  - Four joint models were presented. The best of them scored only 0.5 F_1 points below the winner (Gesmundo et al., 2009)
  - Conclusions with joint models are similar to those obtained in 2008
- No further insights on the two fundamental research questions
- A lot of analysis can still be done on the competition materials. Datasets (available through LDC soon), systems' outputs, etc. represent a very valuable multilingual resource for the future research

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task: Results and Conclusions

- System Architectures
  - Best systems are still pipelined (syntax, then semantics)
  - Four joint models were presented. The best of them scored only 0.5 F_1 points below the winner (Gesmundo et al., 2009)
  - Conclusions with joint models are similar to those obtained in 2008
- No further insights on the two fundamental research questions
- A lot of analysis can still be done on the competition materials. Datasets (available through LDC soon), systems' outputs, etc. represent a very valuable multilingual resource for the future research

Empirical Evaluation: Recent CoNLL Shared Tasks

CoNLL-2009 shared task: Results and Conclusions

- System Architectures
  - Best systems are still pipelined (syntax, then semantics)
  - Four joint models were presented. The best of them scored only 0.5 F_1 points below the winner (Gesmundo et al., 2009)
  - Conclusions with joint models are similar to those obtained in 2008
- No further insights on the two fundamental research questions
- A lot of analysis can still be done on the competition materials. Datasets (available through LDC soon), systems' outputs, etc. represent a very valuable multilingual resource for the future research

Tutorial Overview

1. Introduction
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
   - Generalization to new Domains
   - Dependence on Syntax
   - SRL systems in applications
5. Conclusions

Tutorial Overview

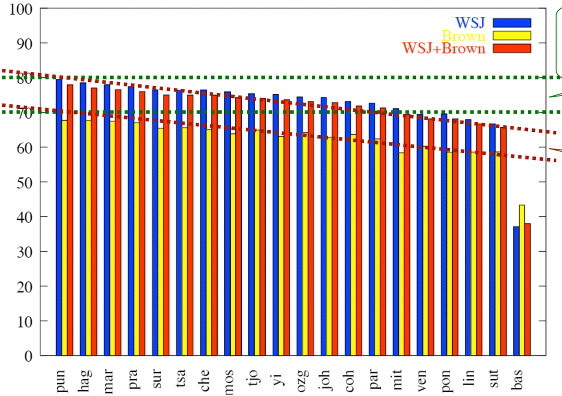
1. Introduction
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
   - Generalization to new Domains
   - Dependence on Syntax
   - SRL systems in applications
5. Conclusions

Domain Dependence

- All statistical ML systems suffer from domain dependence
- How large is this dependence in the case of SRL?
- CoNLL-2005 evaluation: out-of-domain test corpus (Brown) $\Longrightarrow \sim$10 F_1 point drop in performance
- Similar evaluations at CoNLL-2008/2009 shared tasks

Domain Dependence: CoNLL-2005

Results on WSJ and Brown Tests



F_1: 70% ~ 80%
Small differences

Every system suffers from cross-domain test (~10%)

Domain Dependence

Reasons for the low generalization ability

- Training corpus is not representative and big enough (and it will never be)
- The linguistic processors experiment a similar drop in performance
- The loss in accuracy takes place in assigning the semantic roles, not in identification — semantic explanation (Pradhan et al., 2008)

Domain Dependence

Generalization of Role Sets

- Does PropBank numbered core roles allow to generalize across verbs and to unseen predicates in new corpora?
- Aren't thematic role labels (e.g., Agent, Patient, Theme, Experiencer, Source, Beneficiary, etc.) closer to application needs?
- Opportunity: SemLink maps PropBank annotation into VerbNet thematic roles. It covers most of the corpus.
  SL: http://verbs.colorado.edu/semlink/
  VN: http://verbs.colorado.edu/∼mpalmer/projects/verbnet.html

Domain Dependence

Generalization of Role Sets

- Does PropBank numbered core roles allow to generalize across verbs and to unseen predicates in new corpora?
- Aren't thematic role labels (e.g., Agent, Patient, Theme, Experiencer, Source, Beneficiary, etc.) closer to application needs?
- Opportunity: SemLink maps PropBank annotation into VerbNet thematic roles. It covers most of the corpus.
  SL: http://verbs.colorado.edu/semlink/
  VN: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

Domain Dependence

Generalization of Role Sets

- Loper et al. (2007) show that Arg2 generalizes better (in Brown) when training the system from a VerbNet mapped version of PropBank.
- Zapirain et al. (2008) show a negative result:
  - Training on PropBank arguments is more robust under several training settings
  - Also, it is more productive to train on the PropBank roleset and (naively) mapping the output into VerbNet roles, than doing all the process using the VerbNet version of PropBank
- More related studies will be presented at ACL-IJCNLP 2009

Domain Dependence

Generalization of Role Sets

- Loper et al. (2007) show that Arg2 generalizes better (in Brown) when training the system from a VerbNet mapped version of PropBank.
- Zapirain et al. (2008) show a negative result:
  - Training on PropBank arguments is more robust under several training settings
  - Also, it is more productive to train on the PropBank roleset and (naively) mapping the output into VerbNet roles, than doing all the process using the VerbNet version of PropBank
- More related studies will be presented at ACL-IJCNLP 2009

Domain Dependence

Generalization of Role Sets

- Loper et al. (2007) show that Arg2 generalizes better (in Brown) when training the system from a VerbNet mapped version of PropBank.
- Zapirain et al. (2008) show a negative result:
  - Training on PropBank arguments is more robust under several training settings
  - Also, it is more productive to train on the PropBank roleset and (naively) mapping the output into VerbNet roles, than doing all the process using the VerbNet version of PropBank
- More related studies will be presented at ACL-IJCNLP 2009

Domain Dependence

New articles at ACL-IJCNLP 2009

- Merlo & van der Plaas: *Abstraction and Generalisation in Semantic Role Labels: Propbank, VerbNet or both?*
    - Critizism on the experimental settings of (Loper et al. 2007) and (Zapirain et al. 2008): task-oriented evaluation (SRL systems); syntax based; skewed distributions of role labels
    - In the new paper authors analyze how good the two schemes are at capturing the linguistic generalizations that are known to hold for semantic role labels
    - Analyses and statistical measures avoid using syntactic properties or parsing techniques
    - Conclusions: VerbNet is more verb specific and better able to generalize to new semantic role instances; PropBank better capture structural constraints among roles
- Matsubayashi et al.: *A Comparative Study on Generalization of Semantic Roles in FrameNet*

Domain Dependence

New articles at ACL-IJCNLP 2009

- Merlo & van der Plaas: *Abstraction and Generalisation in Semantic Role Labels: Propbank, VerbNet or both?*
  - Critizism on the experimental settings of (Loper et al. 2007) and (Zapirain et al. 2008): task-oriented evaluation (SRL systems); syntax based; skewed distributions of role labels
  - In the new paper authors analyze how good the two schemes are at capturing the linguistic generalizations that are known to hold for semantic role labels
  - Analyses and statistical measures avoid using syntactic properties or parsing techniques
  - Conclusions: VerbNet is more verb specific and better able to generalize to new semantic role instances; PropBank better capture structural constraints among roles
- Matsubayashi et al.: *A Comparative Study on Generalization of Semantic Roles in FrameNet*

Domain Dependence

New articles at ACL-IJCNLP 2009

- Merlo & van der Plaas: *Abstraction and Generalisation in Semantic Role Labels: Propbank, VerbNet or both?*
  - Critizism on the experimental settings of (Loper et al. 2007) and (Zapirain et al. 2008): task-oriented evaluation (SRL systems); syntax based; skewed distributions of role labels
  - In the new paper authors analyze how good the two schemes are at capturing the linguistic generalizations that are known to hold for semantic role labels
  - Analyses and statistical measures avoid using syntactic properties or parsing techniques
  - Conclusions: VerbNet is more verb specific and better able to generalize to new semantic role instances; PropBank better capture structural constraints among roles
- Matsubayashi et al.: *A Comparative Study on Generalization of Semantic Roles in FrameNet*

Domain Dependence

New articles at ACL-IJCNLP 2009

- Merlo & van der Plaas: *Abstraction and Generalisation in Semantic Role Labels: Propbank, VerbNet or both?*
  - Critizism on the experimental settings of (Loper et al. 2007) and (Zapirain et al. 2008): task-oriented evaluation (SRL systems); syntax based; skewed distributions of role labels
  - In the new paper authors analyze how good the two schemes are at capturing the linguistic generalizations that are known to hold for semantic role labels
  - Analyses and statistical measures avoid using syntactic properties or parsing techniques
  - Conclusions: VerbNet is more verb specific and better able to generalize to new semantic role instances; PropBank better capture structural constraints among roles
- Matsubayashi et al.: *A Comparative Study on Generalization of Semantic Roles in FrameNet*

Domain Dependence

Semantic features for SRL

- Motivation
  - Up to now: preeminence of syntactic information in SRL systems
  - Semantic information comes from the raw lexical features
  - But lexical features are sparse and generalize badly to new corpora
- Some works explore the incorporation of selectional preferences as a way to generalize lexical features and gain semantic coherence in the predicate argument structure (Zapirain et al., 2007;2009; Erk, 2007)
- Not easy: a key problem is the noise introduced by lexical ambiguity

Domain Dependence

Semantic features for SRL

- Motivation
  - Up to now: preeminence of syntactic information in SRL systems
  - Semantic information comes from the raw lexical features
  - But lexical features are sparse and generalize badly to new corpora
- Some works explore the incorporation of selectional preferences as a way to generalize lexical features and gain semantic coherence in the predicate argument structure (Zapirain et al., 2007;2009; Erk, 2007)
- Not easy: a key problem is the noise introduced by lexical ambiguity

Domain Dependence

Semantic features for SRL

- Zapirain et al., (2009)
  - Study the use of automatically acquired selectional preferences (SP) for argument classification
  - Setting: *given a verb occurrence and a constituent head word dependant on that verb, assign the most plausible role to the head word according to the selectional preference model*
  - Distributional SP models vs. WordNet-based
  - Lexical features have a high precision but very low recall
  - SP features improve over the baseline: 17 F_1 points on the WSJ datasets and 41 F_1 points on the Brown
  - SP features help to alleviate the lexical sparseness problem
- Initial experiments show significant improvements in a full fledged SRL system (ongoing work)

Domain Dependence

Semantic features for SRL

- Zapirain et al., (2009)
  - Study the use of automatically acquired selectional preferences (SP) for argument classification
  - Setting: *given a verb occurrence and a constituent head word dependant on that verb, assign the most plausible role to the head word according to the selectional preference model*
  - Distributional SP models vs. WordNet-based
  - Lexical features have a high precision but very low recall
  - SP features improve over the baseline: 17 F_1 points on the WSJ datasets and 41 F_1 points on the Brown
  - SP features help to alleviate the lexical sparseness problem
- Initial experiments show significant improvements in a full fledged SRL system (ongoing work)

Domain Dependence

Semantic features for SRL

- Zapirain et al., (2009)
  - Study the use of automatically acquired selectional preferences (SP) for argument classification
  - Setting: *given a verb occurrence and a constituent head word dependant on that verb, assign the most plausible role to the head word according to the selectional preference model*
  - Distributional SP models vs. WordNet-based
  - Lexical features have a high precision but very low recall
  - SP features improve over the baseline: 17 F_1 points on the WSJ datasets and 41 F_1 points on the Brown
  - SP features help to alleviate the lexical sparseness problem
- Initial experiments show significant improvements in a full fledged SRL system (ongoing work)

Tutorial Overview

1. Introduction
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
   - Generalization to new Domains
   - Dependence on Syntax
   - SRL systems in applications
5. Conclusions

Impact of Syntactic Processing in SRL

- SRL results strongly depend on syntax (bottleneck)
- Gold vs. automatic parses: $\sim$90% vs. $\sim$80% F_1
- Drop in performance occurs in identifying argument boundaries

Impact of Syntactic Processing in SRL

- SRL results strongly depend on syntax (bottleneck)
- Gold vs. automatic parses: $\sim$90% vs. $\sim$80% F_1
- Drop in performance occurs in identifying argument boundaries

Impact of Syntactic Processing in SRL

- SRL results strongly depend on syntax (bottleneck)
- Gold vs. automatic parses: $\sim$90% vs. $\sim$80% F_1
- Drop in performance occurs in identifying argument boundaries

Impact of Syntactic Processing in SRL

Partial vs. full parsing (CoNLL-2004/2005)

- Motivation: partial parsing can be more robust to changing application domains
- CoNLL-2005 vs. CoNLL-2004: ~80% vs. ~70% F_1
- ...but the corpus size was the main factor
- The real performance drop when using partial parsing (base chunks + clause boundaries) is ~2 F_1 points (Surdeanu et al., 2007; Punyakanok et al., 2008)
- Bad news: partial parsers degraded their performance as much as full parsers when applied to Brown

Impact of Syntactic Processing in SRL

Partial vs. full parsing (CoNLL-2004/2005)

- Motivation: partial parsing can be more robust to changing application domains
- CoNLL-2005 vs. CoNLL-2004: $\sim$80% vs. $\sim$70% F_1
- ...but the corpus size was the main factor
- The real performance drop when using partial parsing (base chunks + clause boundaries) is $\sim$2 F_1 points (Surdeanu et al., 2007; Punyakanok et al., 2008)
- Bad news: partial parsers degraded their performance as much as full parsers when applied to Brown

Impact of Syntactic Processing in SRL

Partial vs. full parsing (CoNLL-2004/2005)

- Motivation: partial parsing can be more robust to changing application domains
- CoNLL-2005 vs. CoNLL-2004: ∼80% vs. ∼70% F_1
- ...but the corpus size was the main factor
- The real performance drop when using partial parsing (base chunks + clause boundaries) is ∼2 F_1 points (Surdeanu et al., 2007; Punyakanok et al., 2008)
- Bad news: partial parsers degraded their performance as much as full parsers when applied to Brown

Integration of Syntactic Parsing and SRL

First attempt (Yi & Palmer, 2005)

- Syntactic parser trained to predict argument candidates
- Merge the Penn TreeBank and PropBank to generate training parse trees with enriched labels including semantic arguments
- Independent classification of the arguments predicted by the specialized parser
- Results did not improve the conventional architecture
- Possible explanations: weaker base parser / increase in the number of syntactic labels to predict

Integration of Syntactic Parsing and SRL

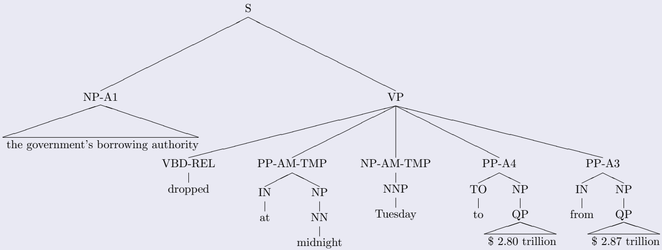
First attempt (Yi & Palmer, 2005)

- Syntactic parser trained to predict argument candidates
- Merge the Penn TreeBank and PropBank to generate training parse trees with enriched labels including semantic arguments
- Independent classification of the arguments predicted by the specialized parser
- Results did not improve the conventional architecture
- Possible explanations: weaker base parser / increase in the number of syntactic labels to predict

Integration of Syntactic Parsing and SRL

Semantic Parsing (Merlo & Musillo, 2008)

- Enrich the annotation of training syntactic trees with semantic role labels



Integration of Syntactic Parsing and SRL

Semantic Parsing (Merlo & Musillo, 2008)

- Train a state-of-the-art parser to produce this new kind of structures (Titov & Henderson, 2007)
- Devise procedures (rule and ML–based) for extracting predicate-argument structures from the enriched trees
- Evaluation on the CoNLL-2005 datasets shows very high precision results (at the price of a low recall)
- Once combined with the best system at CoNLL-2005 the results raise to 80.5% precision, 81.4% recall, and 81.0 F_1-measure for section 23.

Integration of Syntactic Parsing and SRL

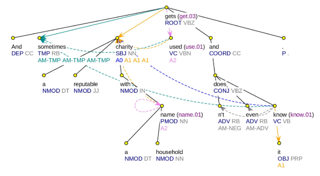
Semantic Parsing (Merlo & Musillo, 2008)

- Train a state-of-the-art parser to produce this new kind of structures (Titov & Henderson, 2007)
- Devise procedures (rule and ML–based) for extracting predicate-argument structures from the enriched trees
- Evaluation on the CoNLL-2005 datasets shows very high precision results (at the price of a low recall)
- Once combined with the best system at CoNLL-2005 the results raise to 80.5% precision, 81.4% recall, and 81.0 F_1-measure for section 23.

Integration of Syntactic Parsing and SRL

Syntactic and Semantic Dependencies (CoNLL-2008/2009)

- A key difficulty: the joint structure is not a dependency tree anymore (Directed Graph); Traditional dependency parsing algorithms work on dependency trees

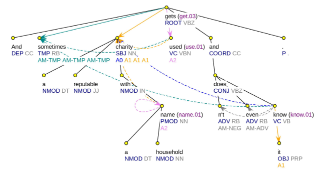


- Three different approaches (from simple to complex)
  1. (Morante et al., 2009)
  2. (Lluís & Màrquez, 2008; Lluís et al., 2009)
  3. (Henderson et al, 2008; Gesmundo et al., 2009)

Integration of Syntactic Parsing and SRL

Syntactic and Semantic Dependencies (CoNLL-2008/2009)

- A key difficulty: the joint structure is not a dependency tree anymore (Directed Graph); Traditional dependency parsing algorithms work on dependency trees



- Three different approaches (from simple to complex)
  1. (Morante et al., 2009)
  2. (Lluís & Màrquez, 2008; Lluís et al., 2009)
  3. (Henderson et al, 2008; Gesmundo et al., 2009)

Integration of Syntactic Parsing and SRL

Approach 1 (Morante et al., 2009)

- Forget about difficult structures and work at word level
- Word classification with extended syntactic-semantic labels

N	Token	Merged Dependencies
1	Housing	2::NMOD:A1
2	starts	2::_:A2 3::SBJ:_ 4::_:A1 6::_:A1 13::_:A0
3	are	0::ROOT:_
4	expected	3::VC:_
5	to	4::OPRD:C-A1
6	quicken	5::IM:_
7	a	8::NMOD:_
8	bit	6::OBJ:A2
9	from	6::ADV:A3
10	August	13::NMOD:AM-TMP
11	's	10::SUFFIX:_
12	annual	13::NMOD:AM-TMP
13	pace	9::PMOD:_
14	of	13::NMOD:A2
15	1,350,000	16::NMOD:_
16	units	14::PMOD:_
17	.	3::P:_

Integration of Syntactic Parsing and SRL

Approach 1 (Morante et al., 2009)

- Three different granularities are considered for class labels (i.e., three overlapping classification problems are defined)
- Make use of Memory Based Learning (insensitivity to large number of classes)
- Add a second layer to construct the structured solution based on the predictions of all word-level classifiers (ranking–based)
- (still) low results at CoNLL-2009 shared task
- Possible reasons: features, heuristics to construct solution, large number of classes, etc.

Integration of Syntactic Parsing and SRL

Approach 1 (Morante et al., 2009)

- Three different granularities are considered for class labels (i.e., three overlapping classification problems are defined)
- Make use of Memory Based Learning (insensitivity to large number of classes)
- Add a second layer to construct the structured solution based on the predictions of all word-level classifiers (ranking–based)
- (still) low results at CoNLL-2009 shared task
- Possible reasons: features, heuristics to construct solution, large number of classes, etc.

Integration of Syntactic Parsing and SRL

Approach 1 (Morante et al., 2009)

- Three different granularities are considered for class labels (i.e., three overlapping classification problems are defined)
- Make use of Memory Based Learning (insensitivity to large number of classes)
- Add a second layer to construct the structured solution based on the predictions of all word-level classifiers (ranking–based)
- (still) low results at CoNLL-2009 shared task
- Possible reasons: features, heuristics to construct solution, large number of classes, etc.

Integration of Syntactic Parsing and SRL

Approach 1 (Morante et al., 2009)

- Three different granularities are considered for class labels (i.e., three overlapping classification problems are defined)
- Make use of Memory Based Learning (insensitivity to large number of classes)
- Add a second layer to construct the structured solution based on the predictions of all word-level classifiers (ranking–based)
- (still) low results at CoNLL-2009 shared task
- Possible reasons: features, heuristics to construct solution, large number of classes, etc.

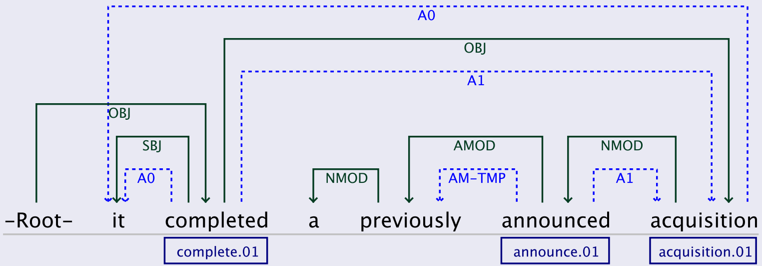
Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)

- Force semantic information to be learnt with the syntactic dependency tree
- Extend regular syntactic dependency parsing algorithms:
  - Minimum Spanning Tree family
  - Eisner algorithm
  - Trained with structure perceptron

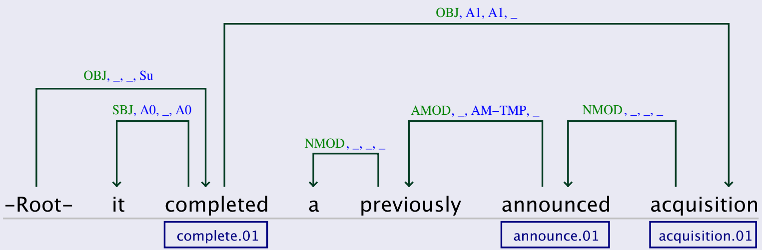
Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)



Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)



Integration of Syntactic Parsing and SRL

Eisner's First Order Dependency Parsing Algorithm

Dependency $d = \langle h, m, l \rangle$

best_tree$(x) = \text{argmax}_{y \in \mathcal{Y}(x)}$ score_tree(y, x)

score_tree$(y, x) = \sum_{\langle h,m,l \rangle \in y}$ score $(\langle h, m, l \rangle, x)$

score$(\langle h, m, l \rangle, x) = \phi(\langle h, m, l \rangle, x) \cdot \mathbf{w}^l$

where

x is an input sentence
y is a dependency tree
$\mathcal{Y}(x)$ is the set of all dependency trees for input x
ϕ is a feature extraction function
$\mathbf{w}^l$ is the weight vector for dependency label l

Integration of Syntactic Parsing and SRL

Eisner's First Order Dependency Parsing Algorithm

Dependency $d = \langle h, m, l \rangle$

best_tree$(x) = \text{argmax}_{y \in \mathcal{Y}(x)}$ score_tree(y, x)

score_tree$(y, x) = \sum_{\langle h,m,l \rangle \in y}$ score $(\langle h, m, l \rangle, x)$

score$(\langle h, m, l \rangle, x) = \phi(\langle h, m, l \rangle, x) \cdot \mathbf{w}^l$

where

x is an input sentence
y is a dependency tree
$\mathcal{Y}(x)$ is the set of all dependency trees for input x
ϕ is a feature extraction function
$\mathbf{w}^l$ is the weight vector for dependency label l

Integration of Syntactic Parsing and SRL

Eisner's First Order Dependency Parsing Algorithm

Dependency $d = \langle h, m, l \rangle$

best_tree$(x) = \text{argmax}_{y \in \mathcal{Y}(x)} \text{score_tree}(y, x)$

score_tree$(y, x) = \sum_{\langle h,m,l \rangle \in y} \text{score}(\langle h, m, l \rangle, x)$

score$(\langle h, m, l \rangle, x) = \phi(\langle h, m, l \rangle, x) \cdot \mathbf{w}^l$

where

x is an input sentence
y is a dependency tree
$\mathcal{Y}(x)$ is the set of all dependency trees for input x
ϕ is a feature extraction function
$\mathbf{w}^l$ is the weight vector for dependency label l

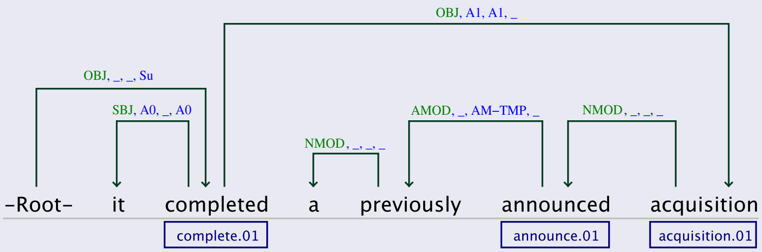
Integration of Syntactic Parsing and SRL

Eisner's First Order Dependency Parsing Algorithm

- The Eisner algorithm is a dynamic programming search algorithm that computes the best first-order factorized tree in $O(n^3)$ (i.e., solves the argmax function).
- All binary linear classifiers can be trained on-line using structure preceptron (Collins & Duffy 2001; Carreras et al., 2007;2008)
- Can be naturally extended to higher order factorizations, e.g., (Carreras, 2007)

Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)



Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)

An extended dependency is:

$$d = \left\langle h, m, l_{syn}, l_{sem\ p_1}, \ldots, l_{sem\ p_q} \right\rangle$$

h is the head
m the modifier
l_{syn} the syntactic label
$l_{sem\ p_i}$ one semantic label for each sentence predicate p_i

Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)

$$\text{best\_tree}(x, y') = \text{argmax}_{y \in \mathcal{Y}(x)} \text{score\_tree}(y, x, y')$$

$$\text{score\_tree}(y, x, y') = \sum_{\langle h,m,l_{syn},\mathbf{l}\rangle \in y} \text{score}(\langle h, m, l_{syn}, \mathbf{l}\rangle, x, y')$$

$$\text{score}(\langle h, m, l_{syn}, \mathbf{l}\rangle, x, y') = \text{synt\_score}(\langle h, m, l_{syn}\rangle, x) + \text{sem\_score}(\langle h, m, \mathbf{l}\rangle, x, y')$$

$\mathbf{l} = l_{sem\ p_1}, \ldots, l_{sem\ p_q}$ are the semantic labels for predicates p_i

$$\text{sem\_score}(h, m, l_{sem\ p_1}, \ldots, l_{sem\ p_q}, x, y') =$$

$$\sum_{l_{sem\ p_i}} \frac{\phi_{sem}(\langle h, m, l_{sem\ p_i}\rangle, p_i, x, y') \cdot \mathbf{w}^{(l_{sem\ p_i})}}{q}$$

Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)

$$\text{best\_tree}(x, y') = \text{argmax}_{y \in \mathcal{Y}(x)} \text{score\_tree}(y, x, y')$$

$$\text{score\_tree}(y, x, y') = \sum_{\langle h,m,l_{syn},\mathbf{l}\rangle \in y} \text{score}(\langle h, m, l_{syn}, \mathbf{l}\rangle, x, y')$$

$$\text{score}(\langle h, m, l_{syn}, \mathbf{l}\rangle, x, y') = \text{synt\_score}(\langle h, m, l_{syn}\rangle, x) + \text{sem\_score}(\langle h, m, \mathbf{l}\rangle, x, y')$$

$\mathbf{l} = l_{sem\ p_1}, \ldots, l_{sem\ p_q}$ are the semantic labels for predicates p_i

$$\text{sem\_score}(h, m, l_{sem\ p_1}, \ldots, l_{sem\ p_q}, x, y') = \sum_{l_{sem\ p_i}} \frac{\phi_{sem}(\langle h, m, l_{sem\ p_i}\rangle, p_i, x, y') \cdot \mathbf{w}^{(l_{sem\ p_i})}}{q}$$

Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)

- Eisner inference unchanged (the only change occurs at dependency scoring)
- Standard syntactic and SRL features
- On-line training of **w** vectors using structure perceptron
- Extension to second-order parsing is straightforward
- Moderate results at CoNLL-2008 and 2009 shared tasks
- Difficulties: 1) too complex decisions at dependency level (semantic structure is not exploited); 2) adjustment of the relative weight of syntactic and semantic contributions

Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)

- Eisner inference unchanged (the only change occurs at dependency scoring)
- Standard syntactic and SRL features
- On-line training of **w** vectors using structure perceptron
- Extension to second-order parsing is straightforward
- Moderate results at CoNLL-2008 and 2009 shared tasks
- Difficulties: 1) too complex decisions at dependency level (semantic structure is not exploited); 2) adjustment of the relative weight of syntactic and semantic contributions

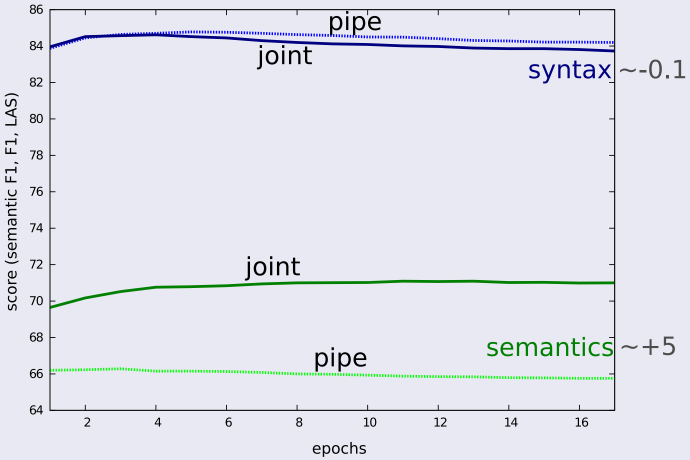
Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)

- Eisner inference unchanged (the only change occurs at dependency scoring)
- Standard syntactic and SRL features
- On-line training of **w** vectors using structure perceptron
- Extension to second-order parsing is straightforward
- Moderate results at CoNLL-2008 and 2009 shared tasks
- Difficulties: 1) too complex decisions at dependency level (semantic structure is not exploited); 2) adjustment of the relative weight of syntactic and semantic contributions

Integration of Syntactic Parsing and SRL

Approach 2 (Lluís & Màrquez, 2008; Lluís et al., 2009)



Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

- Deal with syntax and semantics as separate structures
- but synchronize the generation of both structures
- and establish dependencies between both levels in the form of latent variables
- Transition-based model of parsing (*shift-reduce* style or *history-based*)
- New operation (*swap*) for on-line planarisation of the semantic graph
- Synchronous derivations are modeled with an Incremental Sigmoid Belief Network (ISBN; Titov and Henderson's parser, 2007)

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

- Deal with syntax and semantics as separate structures
- but synchronize the generation of both structures
- and establish dependencies between both levels in the form of latent variables
- Transition-based model of parsing (*shift-reduce* style or *history-based*)
- New operation (*swap*) for on-line planarisation of the semantic graph
- Synchronous derivations are modeled with an Incremental Sigmoid Belief Network (ISBN; Titov and Henderson's parser, 2007)

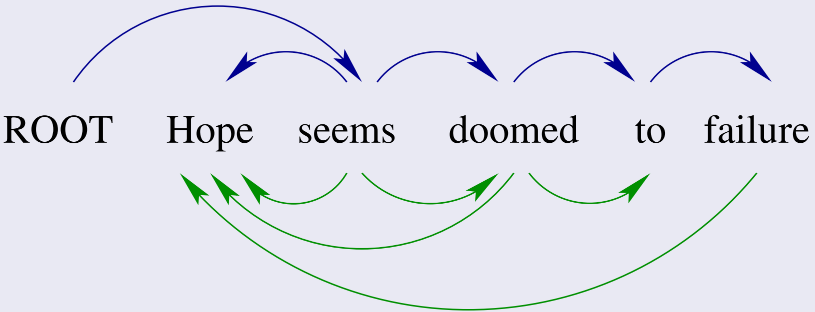
Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

- Deal with syntax and semantics as separate structures
- but synchronize the generation of both structures
- and establish dependencies between both levels in the form of latent variables
- Transition-based model of parsing (*shift-reduce* style or *history-based*)
- New operation (*swap*) for on-line planarisation of the semantic graph
- Synchronous derivations are modeled with an Incremental Sigmoid Belief Network (ISBN; Titov and Henderson's parser, 2007)

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)



$P(T_d, T_s)$

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Define **two separate derivations**, one for the syntactic structure and one for the semantic structure.

$$P(T_d, T_s) = P(D_d^1, ..., D_d^{m_d}, D_s^1, ..., D_s^{m_s})$$

Use an intermediate synchronization granularity, between full predications and individual actions: synchronization at **each word** prediction

$$C^t = D_d^{b_d^t}, ..., D_d^{e_d^t}, shift_t, D_s^{b_s^t}, ..., D_s^{e_s^t}, shift_t$$

$$P(D_d^1, ..., D_d^{m_d}, D_s^1, ..., D_s^{m_s}) = P(C^1, \ldots, C^n)$$

- Results in **one shared input queue**
  Allows **two separate stacks**

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

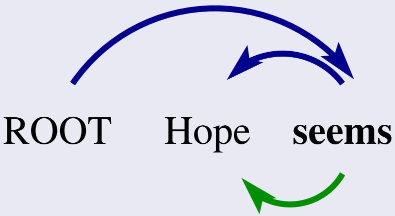
ROOT **Hope**

$\mathbf{P(C^1)}$

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

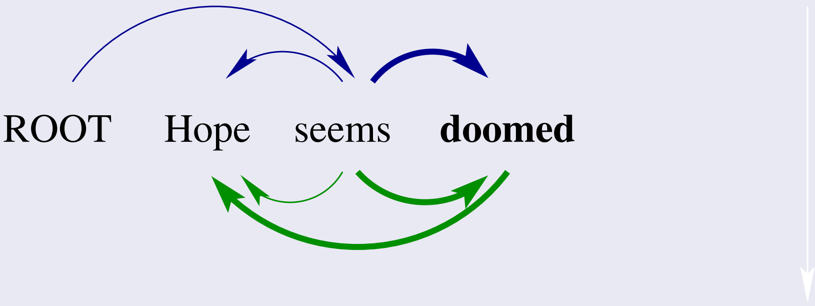


$P(C^1)\ \mathbf{P}(\mathbf{C}^2|\mathbf{C}^1)$

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

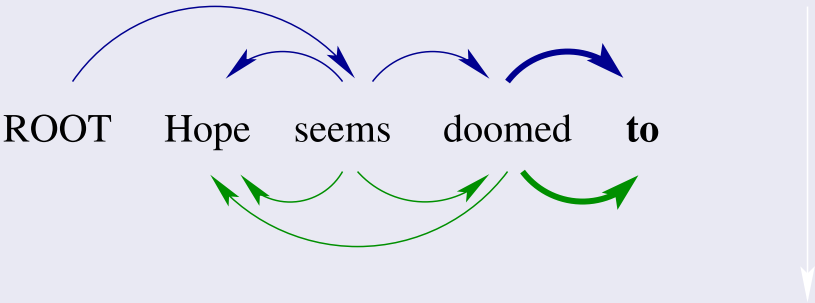


$P(C^1)\ P(C^2|C^1)\ \mathbf{P(C^3|C^1, C^2)}$

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

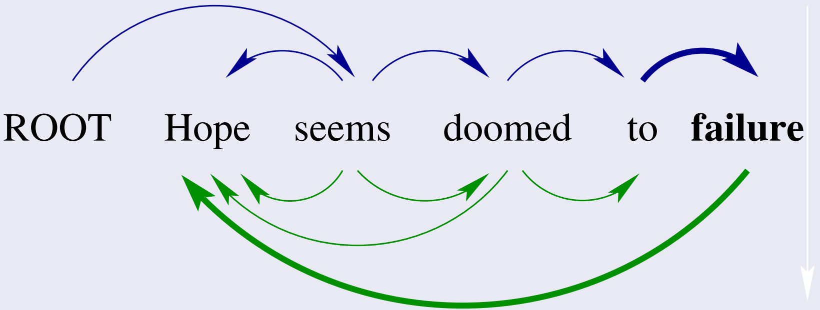


$P(C^1)\ P(C^2|C^1)\ P(C^3|C^1, C^2)\ \mathbf{P(C^4|C^1, C^2, C^3)}$

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)



$$P(C^1)\ P(C^2|C^1)\ P(C^3|C^1,C^2)\ P(C^4|C^1,C^2,C^3)\ \mathbf{P(C^5|C^1,C^2,C^3,C^4)}$$

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

ROOT Hope

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

ROOT Hope seems

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

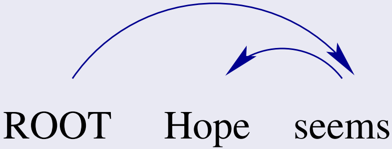
ROOT Hope seems

Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

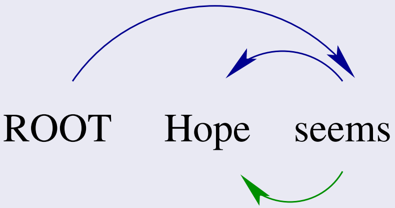


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

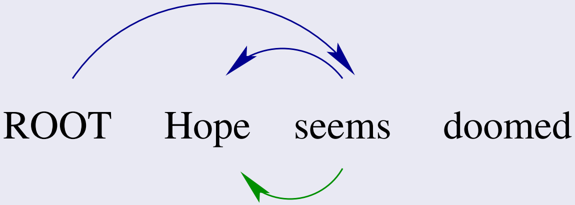


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

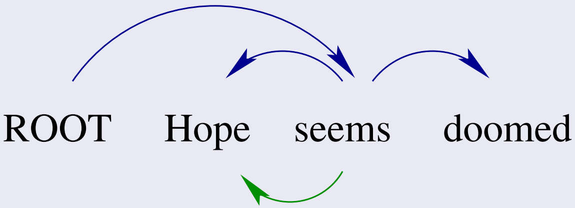


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

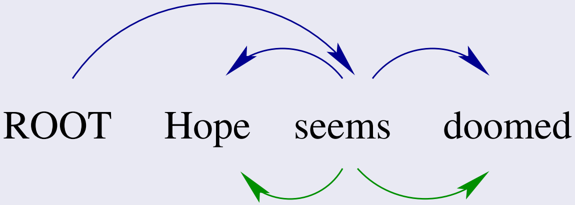


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

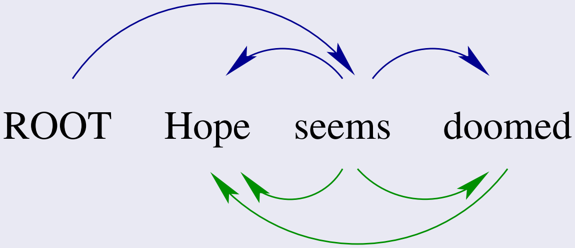


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

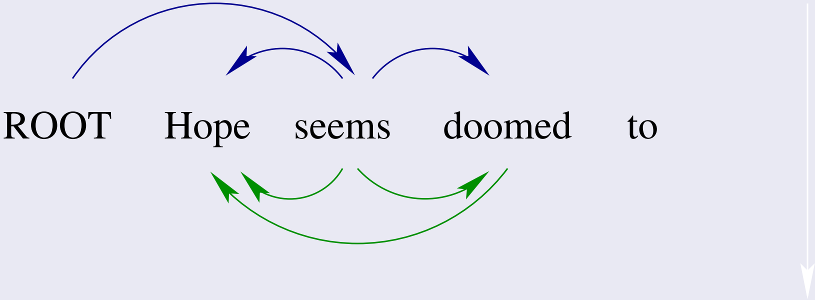


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

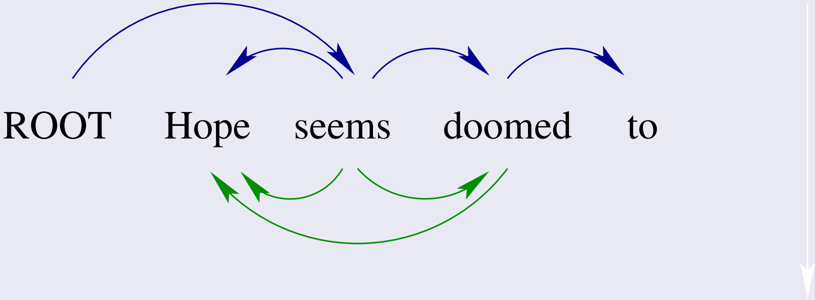


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

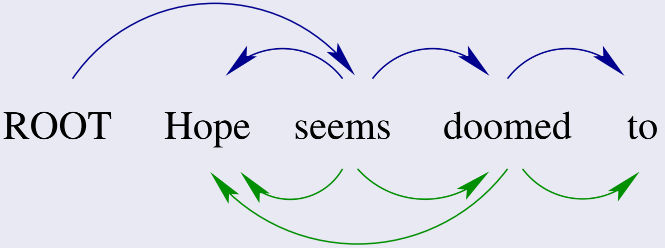


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

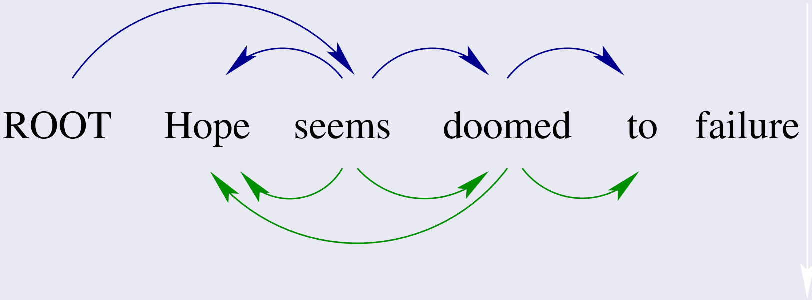


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

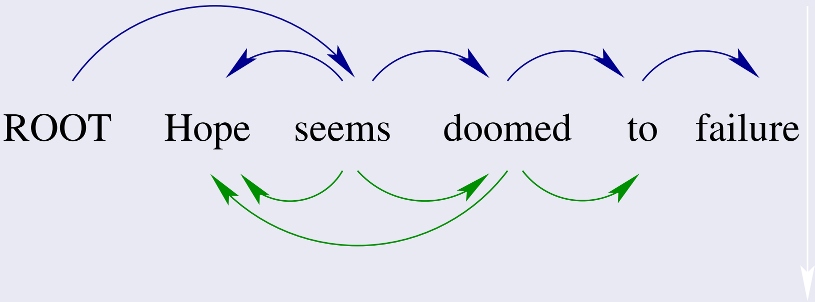


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

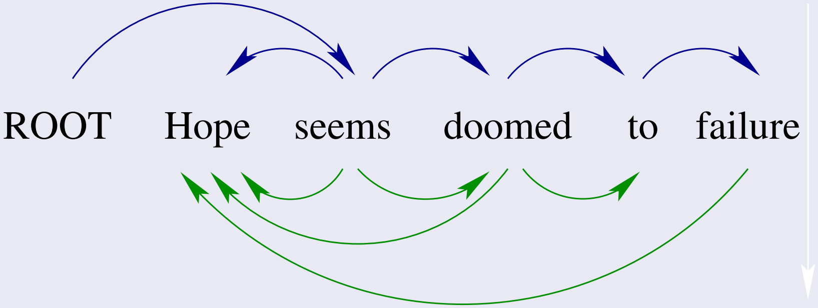


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

Derivation example:

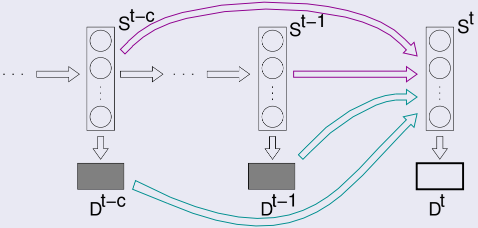


Slides by James Henderson

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

- ISBNs are Dynamic Bayesian Networks **for modeling structures**,
- with **vectors of latent variables** annotating derivation states
- Connections between latent states reflect locality in the syntactic or semantic **structure**,
- Explicit conditioning features of the history are also specified



Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

- The model maximizes the joint probability of the syntactic and semantic dependencies ($\Longrightarrow$ enforces that the output structure be globally coherent)
- Good results at CoNLL-2008: joint parsing improves the semantic part by 3.5 F_1 points
- Very good results at CoNLL-2009: F_1 score 82.14 (3rd position; almost tied with the two first). The parser proved to be very robust across languages and data domains

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

- The model maximizes the joint probability of the syntactic and semantic dependencies ($\Longrightarrow$ enforces that the output structure be globally coherent)
- Good results at CoNLL-2008: joint parsing improves the semantic part by 3.5 F_1 points
- Very good results at CoNLL-2009: F_1 score 82.14 (3rd position; almost tied with the two first). The parser proved to be very robust across languages and data domains

Impact of Syntactic Processing in SRL

Approach 3 (Henderson et al., 2008; Gesmundo et al., 2009)

- The model maximizes the joint probability of the syntactic and semantic dependencies ($\Longrightarrow$ enforces that the output structure be globally coherent)
- Good results at CoNLL-2008: joint parsing improves the semantic part by 3.5 F_1 points
- Very good results at CoNLL-2009: F_1 score 82.14 (3rd position; almost tied with the two first). The parser proved to be very robust across languages and data domains

Tutorial Overview

1. Introduction
2. State-of-the-art
3. Empirical evaluation and lessons learned
4. Problems and challenges
   - Generalization to new Domains
   - Dependence on Syntax
   - SRL systems in applications
5. Conclusions

SRL in Applications

Examples of applications of SRL

- Information Extraction (Surdeanu et al., 2003)
- Question & Answering (Narayanan and Harabagiu, 2004; Frank et al., 2007)
- Automatic Summarization (Melli et al., 2005)
- Coreference Resolution (Ponzetto and Strube, 2006)
- See (Yih & Toutanova, 2006) tutorial for a discussion on all previous works

SRL in Applications

Examples of applications of SRL

- Information Extraction (Surdeanu et al., 2003)
- Question & Answering (Narayanan and Harabagiu, 2004; Frank et al., 2007)
- Automatic Summarization (Melli et al., 2005)
- Coreference Resolution (Ponzetto and Strube, 2006)
- See (Yih & Toutanova, 2006) tutorial for a discussion on all previous works

SRL in Applications

Other applications of SRL

- Machine Translation Evaluation
  (Giménez and Màrquez, 2007)
- Machine Translation
  (Boas, 2002; Wu and Fung, 2009a;2009b)
- Textual Entailment
  (Tatu & Moldovan, 2005; Burchardt et al., 2007)
- Modeling Early Language Acquisition (Connor et al., 2008;2009)
- Pictorial Communication Systems (Goldberg, et al., 2008)
- ...
- We will concentrate on Machine Translation

SRL in Applications

Other applications of SRL

- Machine Translation Evaluation
  (Giménez and Màrquez, 2007)
- Machine Translation
  (Boas, 2002; Wu and Fung, 2009a;2009b)
- Textual Entailment
  (Tatu & Moldovan, 2005; Burchardt et al., 2007)
- Modeling Early Language Acquisition (Connor et al., 2008;2009)
- Pictorial Communication Systems (Goldberg, et al., 2008)
- ...
- We will concentrate on Machine Translation

SRL in Machine Translation

Automatic MT Evaluation

- Giménez and Màrquez (2007;2008)
  - Introduced a new set of automatic metrics for MT evaluation based on rich linguistic information (including similarity at lexical, shallow/deep syntactic, shallow/deep semantic levels)
  - SRL provides an important subset of these features
  - Measuring the overlap of semantic roles between the system's output and reference target sentences helps improving correlation with human judgement of translation quality

SRL in Machine Translation

Automatic MT Evaluation

- Giménez and Màrquez (2007;2008)
  - Introduced a new set of automatic metrics for MT evaluation based on rich linguistic information (including similarity at lexical, shallow/deep syntactic, shallow/deep semantic levels)
  - SRL provides an important subset of these features
  - Measuring the overlap of semantic roles between the system's output and reference target sentences helps improving correlation with human judgement of translation quality

SRL in Machine Translation

Automatic MT Evaluation

- Giménez and Màrquez (2007;2008)
  - Surprinsingly robust SRL parsing for ill-formed sentences (works well for system comparison and ranking)
  - Better than BLEU-like lexical measures, especially in heterogeneous scenarios and out-of-domain evaluation
  - IQMT suite is freely available http://www.lsi.upc.edu/∼nlp/IQMT/

SRL in Machine Translation

Automatic MT Evaluation

- Giménez and Màrquez (2007;2008)
  - Surprinsingly robust SRL parsing for ill-formed sentences (works well for system comparison and ranking)
  - Better than BLEU-like lexical measures, especially in heterogeneous scenarios and out-of-domain evaluation
  - IQMT suite is freely available
    http://www.lsi.upc.edu/∼nlp/IQMT/

SRL in Machine Translation

Exploring the application of SRL in SMT

- Wu and Fung (2009a)

  Present a series of experiments to study the potential impact of SRL in improving MT accuracy. Three basic questions:

  1. Do current SMT systems produce good translations at predicate structure level? Not really (even when the predicate is correctly translated)
  2. Does incorporating SR analysis contribute anything beyond the current work on syntactic SMT models? SR enforce cross-lingual translation patterns more correctly
  3. What is the potential quantitative impact of realistic SR guidance to SMT systems? significant BLEU and METEOR improvement by >2 points

SRL in Machine Translation

Exploring the application of SRL in SMT

- Wu and Fung (2009a)

  Present a series of experiments to study the potential impact of SRL in improving MT accuracy. Three basic questions:

  1. Do current SMT systems produce good translations at predicate structure level? Not really (even when the predicate is correctly translated)
  2. Does incorporating SR analysis contribute anything beyond the current work on syntactic SMT models? SR enforce cross-lingual translation patterns more correctly
  3. What is the potential quantitative impact of realistic SR guidance to SMT systems? significant BLEU and METEOR improvement by >2 points

SRL in Machine Translation

Exploring the application of SRL in SMT

- Wu and Fung (2009a)

  Present a series of experiments to study the potential impact of SRL in improving MT accuracy. Three basic questions:

  1. Do current SMT systems produce good translations at predicate structure level? Not really (even when the predicate is correctly translated)
  2. Does incorporating SR analysis contribute anything beyond the current work on syntactic SMT models? SR enforce cross-lingual translation patterns more correctly
  3. What is the potential quantitative impact of realistic SR guidance to SMT systems? significant BLEU and METEOR improvement by >2 points

SRL in Machine Translation

First SMT system with SRL

- (Wu and Fung, 2009b)
  - Hybrid SMT system incorporating Semantic Role Labeling and phase-based SMTmodels
  - Two-pass architecture: 1) phrase-based SMT system; 2) reordering guided by shallow semantic parsers
  - SRL is performed first into source and output sentences in order to identify predicate structures and constituents to be re-ordered.
  - Then, a set of candidate re-ordered sentences are generated (by moving SR-mismatched constituents)
  - Finally, a SRL parser is applied to the candidate translations and the best match with the input structure is returned

SRL in Machine Translation

First SMT system with SRL

- (Wu and Fung, 2009b)
  - Hybrid SMT system incorporating Semantic Role Labeling and phase-based SMTmodels
  - Two-pass architecture: 1) phrase-based SMT system; 2) reordering guided by shallow semantic parsers
  - SRL is performed first into source and output sentences in order to identify predicate structures and constituents to be re-ordered.
  - Then, a set of candidate re-ordered sentences are generated (by moving SR-mismatched constituents)
  - Finally, a SRL parser is applied to the candidate translations and the best match with the input structure is returned

SRL in Machine Translation

First SMT system with SRL

- (Wu and Fung, 2009b)
  - Hybrid SMT system incorporating Semantic Role Labeling and phase-based SMTmodels
  - The hybrid model produces a slight but significant improvement in the quality of the translations (measured with BLEU score)
  - Chinese-English translation on Newswire texts

Tutorial Overview

1 Introduction

2 State-of-the-art

3 Empirical evaluation and lessons learned

4 Problems and challenges

5 Conclusions

General Conclusions

- SRL is an important problem in NLP with strong connections to applications requiring some degree of semantic interpretation
- It is a very active topic of research, which has generated an important body of work in the last 6 years
- Some news are good but...
- SRL still has to face important challenges before we see systems in real open-domain applications
- Good opportunities for future research on the topic

General Conclusions

- SRL is an important problem in NLP with strong connections to applications requiring some degree of semantic interpretation
- It is a very active topic of research, which has generated an important body of work in the last 6 years
- Some news are good but...
- SRL still has to face important challenges before we see systems in real open-domain applications
- Good opportunities for future research on the topic

General Conclusions

- SRL is an important problem in NLP with strong connections to applications requiring some degree of semantic interpretation
- It is a very active topic of research, which has generated an important body of work in the last 6 years
- Some news are good but...
- SRL still has to face important challenges before we see systems in real open-domain applications
- Good opportunities for future research on the topic

Specific Conclusions

- Generalization to new events/domains/corpora is a very weak point of statistical SRL systems
  - System portability must be improved (e.g., domain adaptation, appropriate role sets, lexical semantic generalization, etc.)
- System complexity is increasing in a higher scale than performance
  - SRL systems have to be more efficient for massive text processing

Specific Conclusions

- Generalization to new events/domains/corpora is a very weak point of statistical SRL systems
  - System portability must be improved (e.g., domain adaptation, appropriate role sets, lexical semantic generalization, etc.)
- System complexity is increasing in a higher scale than performance
  - SRL systems have to be more efficient for massive text processing

Specific Conclusions

- SRL Systems for languages other than English should be developed and made available to the NLP community
- Reduce the cost of producing semantically annotated corpora for under resourced languages (e.g., making use of semi-supervised training, corpora in other languages, etc.)

Specific Conclusions

- SRL technology should provide significant improvements in widely used NLP applications. A jump is needed from the laboratory conditions to the real world.
- Investigate learning architectures that take advantage of the joint resolution of several syntactic–semantic levels (parsing, SRL, WSD, NEs, coreference, etc.)

Specific Conclusions

- SRL technology should provide significant improvements in widely used NLP applications. A jump is needed from the laboratory conditions to the real world.
- Investigate learning architectures that take advantage of the joint resolution of several syntactic–semantic levels (parsing, SRL, WSD, NEs, coreference, etc.)

Acknowledgements

Thanks to

- ACL–IJCNLP 2009 Organizers and Tutorial Chairs (Diana McCarthy and Chengqing Zong)
- All people that directly or indirectly helped me with the materials presented in this tutorial: Mihai Surdeanu, Xavier Carreras, Xavier Lluís, Dan Roth, Kristina Toutanova, Wen-Tau Yih, Jan Hajič, Jesús Giménez, Paola Merlo and James Henderson
- Spanish Ministry of Education and Science for the partial funding of author's research (OpenMT, TIN2006-15307-C03-01)
- Last, but not least, thanks to all tutorial attendees!

Semantic Role Labeling
Past, Present and Future

Lluís Màrquez

TALP Research Center

Tecnhical University of Catalonia

Tutorial at ACL-IJCNLP 2009

Suntec – Singapore

August 2, 2009

—Version from August 3, 2009—