Exploring challenges in Semantic Role Labeling

Lluís Màrquez
TALP Research Center
Tecnhical University of Catalonia

Invited talk at ABBYY Open Seminar
Moscow, Russia, May 28, 2013

Talk Overview

The Problem

Semantic Role Labeling

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

The Problem

IE point of view

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

[The luxury auto maker]$_{AGENT}$ [last year]$_{TEMP}$ sold$_{P}$ [1,214 cars]$_{OBJECT}$ [in the U.S.]$_{LOCATIVE}$

SRL can be very useful for many practical NLP applications: IE, Q&A, Machine Translation, Summarization, etc.

The Problem

IE point of view

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

[The luxury auto maker]$_{AGENT}$ [last year]$_{TEMP}$ sold$_{P}$ [1,214 cars]$_{OBJECT}$ [in the U.S.]$_{LOCATIVE}$

SRL can be very useful for many practical NLP applications: IE, Q&A, Machine Translation, Summarization, etc.

The Problem

IE point of view

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

[The luxury auto maker]$_{AGENT}$ [last year]$_{TEMP}$ sold$_{P}$ [1,214 cars]$_{OBJECT}$ [in the U.S.]$_{LOCATIVE}$

SRL can be very useful for many practical NLP applications: IE, Q&A, Machine Translation, Summarization, etc.

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
- Yesterday Scott was hit by Kristina with a baseball
- Kristina hit Scott with a baseball yesterday

$\Rightarrow$ All of them share the same semantic representation:

hit(Kristina,Scott,yesterday,with a baseball)

Example from (Yih & Toutanova, 2006)

The Problem

Syntactic variations

$\underbrace{\text{Yesterday}}^{\text{TEMP}}$, $\overbrace{\text{Kristina}}^{\text{HITTER}}$ 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
- Yesterday Scott was hit by Kristina with a baseball
- Kristina hit Scott with a baseball yesterday

⇒ All of them share the same semantic representation:

hit(Kristina,Scott,yesterday,with a baseball)

Example from (Yih & Toutanova, 2006)

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$_{B-RECIP}$ me$_{I}$ this$_{B-TMP}$ morning$_{I}$.$_{O}$

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$_{B-RECIP}$ me$_{I}$ this$_{B-TMP}$ morning$_{I}$.$_{O}$

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

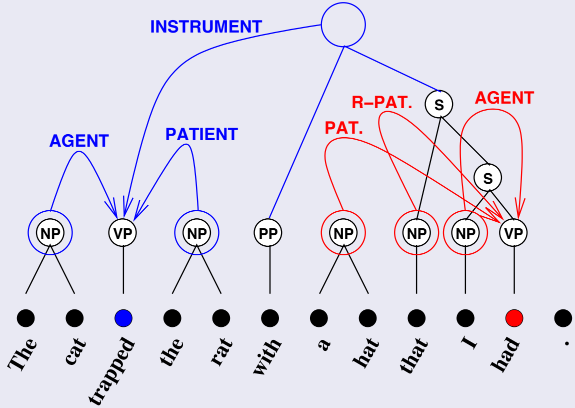
" Mr. Smith sent the report to me this morning . "

[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$_{B-RECIP}$ me$_{I}$ this$_{B-TMP}$ morning$_{I}$.$_{O}$

The Problem

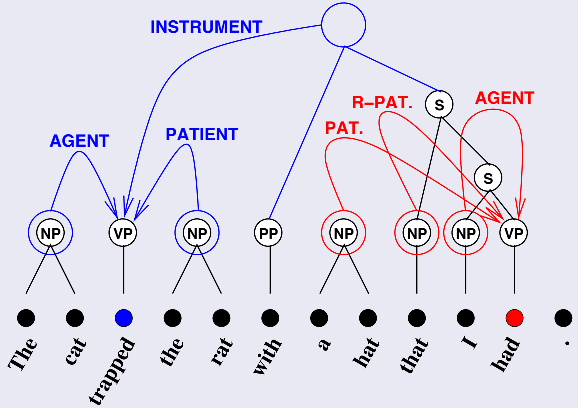
Structural View



Output is a *hierarchy of labeled arguments*

The Problem

Structural View

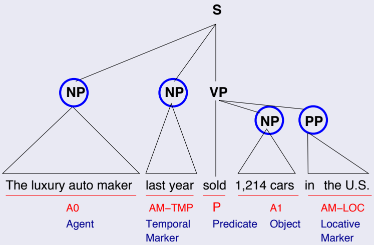


Output is a *hierarchy of labeled arguments*

The Problem

Linguistic nature of the problem

- Argument identification is strongly related to syntax



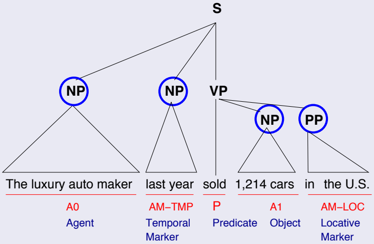
- Role labeling is a semantic task

  (e.g., selectional preferences could play an important role)

The Problem

Linguistic nature of the problem

- Argument identification is strongly related to syntax



- Role labeling is a semantic task

  (e.g., selectional preferences could play an important role)

The Statistical Approach to SRL

A pipeline architecture in 3 steps

1. *Select argument candidates*

   ⇒ Parse the sentence and apply heuristics to select a compact subset of syntactic constituents as candidate arguments

2. *Local scoring of candidates*

   ⇒ Apply classifiers locally to candidate arguments to identify actual arguments and label them with semantic roles

3. *Joint scoring of complete solutions*

   ⇒ Apply inference to enforce globally good predicate-argument structures (ILP, re-ranking, structure learning, etc.)

The Statistical Approach to SRL

A pipeline architecture in 3 steps

1. *Select argument candidates*

   ⇒ Parse the sentence and apply heuristics to select a compact subset of syntactic constituents as candidate arguments

2. *Local scoring of candidates*

   ⇒ Apply classifiers locally to candidate arguments to identify actual arguments and label them with semantic roles

3. *Joint scoring of complete solutions*

   ⇒ Apply inference to enforce globally good predicate-argument structures (ILP, re-ranking, structure learning, etc.)

The Statistical Approach to SRL

A pipeline architecture in 3 steps

1. *Select argument candidates*
   $\Rightarrow$ Parse the sentence and apply heuristics to select a compact subset of syntactic constituents as candidate arguments
2. *Local scoring of candidates*
   $\Rightarrow$ Apply classifiers locally to candidate arguments to identify actual arguments and label them with semantic roles
3. *Joint scoring of complete solutions*
   $\Rightarrow$ Apply inference to enforce globally good predicate-argument structures (ILP, re-ranking, structure learning, etc.)

Talk Overview

Joint work with

Eneko Agirre, *Mihai Surdeanu* and *Beñat Zapirain*

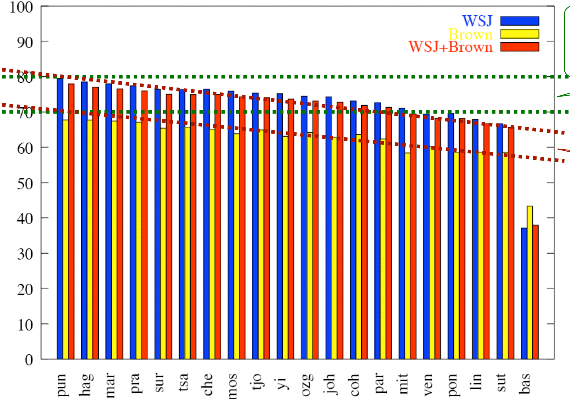
(Zapirain et al. 2010) — ACL

(Zapirain et al. 2011) — NAACL

(Zapirain et al. 2013) — Computational Linguistics 39(3)

Results from CoNLL-2005 shared task

Results on WSJ and Brown Tests



F_1: 70% ~ 80%
Small differences

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

Results from CoNLL-2005 shared task

Reasons for the low generalization ability

- The training corpus is not representative and large enough (and it will never be)
- Taggers and syntactic parsers also experience a significant drop in performance
- The main loss in performance takes place in role classification, not identification — semantic explanation (Pradhan et al., 2008)

Semantic Features for SRL

Motivation

- Most current systems capture semantics through lexicalized features on the predicate and the head word of the argument to be classified
- But lexical features are sparse and generalize badly

  [JFK]$_{Patient}$ *was_assassinated* [in Dallas]$_{LOC}$

  [JFK]$_{Patient}$ *was_assassinated* [in November]$_{TMP}$

- [in Texas]$_{???}$, [in autumn]$_{???}$

Semantic Features for SRL

Motivation

- Most current systems capture semantics through lexicalized features on the predicate and the head word of the argument to be classified
- But lexical features are sparse and generalize badly

  [JFK]$_{Patient}$ *was_assassinated* [in Dallas]$_{LOC}$

  [JFK]$_{Patient}$ *was_assassinated* [in November]$_{TMP}$

- [in Texas]$_{???}$, [in autumn]$_{???}$

Semantic Features for SRL

Motivation

- Most current systems capture semantics through lexicalized features on the predicate and the head word of the argument to be classified
- But lexical features are sparse and generalize badly

  [JFK]$_{Patient}$ *was_assassinated* [in Dallas]$_{LOC}$

  [JFK]$_{Patient}$ *was_assassinated* [in November]$_{TMP}$

- [in Texas]$_{???}$, [in autumn]$_{???}$

Semantic Features for SRL

Motivation

Selectional Preferences and distributional similarity techniques should help us to classify arguments with low–frequency or unknown head words

[Dallas $\approx$ Texas]$_{Location}$, [November $\approx$ autumn]$_{Temporal}$

Previous Work

Selectional Preferences

- Modeling semantic preferences that predicates impose on their arguments
- Long tradition of automatic acquisition of selectional preferences (SPs) from corpora. WordNet–based and distributional models of SPs
  (Resnik, 1993; Pantel and Lin, 2000; Brockmann and Lapata, 2003) (Erk 2007; Erk et al., 2011; etc.)
  - ⇒ e.g., estimate plausibility of triples: (verb, argument, head-word)
  - ⇒ useful for syntactic-semantic disambiguation

Previous Work

SPs applied to Semantic Role Labeling

- (Gildea and Jurafsky, 2002) – FrameNet
  - ⇒ First researchers to apply selectional preferences to SRL
  - ⇒ Distributional clustering and WordNet-based techniques to generalize argument heads
  - ⇒ Slight improvement in role classification (NP arguments)
- Zapirain et al. (2010; 2013) – PropBank
  - ⇒ Show that selectional preferences can improve semantic role classification in a state-of-the-art SRL system

Selectional Preferences for SRL

(Zapirain et al., 2013)

Two types of selectional preferences (SP)

i. verb–*role*: list of heads of NP arguments of the predicate verb that are labeled with the role *role*

```
write-Arg0:  Angrist anyone baker ball bank Barlow Bates ...
write-Arg1:  abstract act analysis article asset bill book ...
write-Arg2:  bank commander hundred jaguar Kemp member ...
write-AM-LOC: paper space ...

...
```

ii. prep–*role*: list of nominal heads of PP arguments with preposition prep that are labeled with the role *role*

```
from-Arg2:  academy account acquisition activity ad ...
from-Arg3:  activity advertising agenda airport ...
from-Arg4:  europe Golenbock system Vizcaya west
from-AM-TMP: april august beginning bell day dec.  half ...
from-AM-LOC: agency area asia body bureau orlando ...

...
```

Selectional Preferences for SRL (Zapirain et al., 2013)

SP models: $SP_{sim}(p, r, w)$ compatibility score

- Discriminative approach: given a new argument of a predicate p, we compare its head (w) to the selectional preference of each possible role label r, i.e., we want to find the role with the selectional preference that fits the head best
- We compute the compatibility scores using two different methods
  - ⇒ WordNet based —using (Resnik, 1993)
  - ⇒ Based on distributional similarity —a la Erk (2007)

Selectional Preferences for SRL (Zapirain et al., 2013)

SP models: $SP_{sim}(p, r, w)$ compatibility score

- Discriminative approach: given a new argument of a predicate p, we compare its head (w) to the selectional preference of each possible role label r, i.e., we want to find the role with the selectional preference that fits the head best
- We compute the compatibility scores using two different methods
  - ⇒ WordNet based —using (Resnik, 1993)
  - ⇒ Based on distributional similarity —a la Erk (2007)

Selectional Preferences for SRL (Zapirain et al., 2013)

WordNet SP models

- Resnik formula (1993) is used to precalculate a weighted list of relevant synsets for the lists of words contained in the SPs

SP write–*Arg0*: Angrist anyone baker ball bank Barlow Bates ...

n#00002086 5.875 life form organism being living thing "any living entity"
n#00001740 5.737 entity something "anything having existence (living or nonliving)"
n#00009457 4.782 object physical object "a physical (tangible and visible) entity;"
n#00004123 4.351 person individual someone somebody mortal human soul "a human being;"
...

SP write–*Arg1*: abstract act analysis article asset bill book ...

n#00019671 7.956 communication "something that is communicated between people or groups"
n#04949838 4.257 message content subject matter substance "what a communication that ..."
n#00018916 3.848 relation "an abstraction belonging to or characteristic of two entities"
n#00013018 3.574 abstraction "a concept formed by extracting common features from examples"
...

Selectional Preferences for SRL (Zapirain et al., 2013)

WordNet SP models

- At test time, for a new argument of the predicate write with head word book:

  ⇒ consider $S = \{$<book>$\} \cup$ “all its hypernyms in WordNet” (for all senses of book)

  ⇒ $SP_{Res}(\mathbf{write}, Arg1, book)$ returns the sum of the weights of the sysnsets in S matching the synsets in the list corresponding to the SP write–*Arg1*

Selectional Preferences for SRL (Zapirain et al., 2013)

Distributional SP models: based on Erk's (2007) setting

JFK was assassinated [in Texas]???

SP in–*TMP*: *November, century, month*

SP in–*LOC*: *Dallas, railway, city*

$$SP_{sim}(p,r,w) = \sum_{w_i \in Seen(p,r)} sim(w,w_i) \cdot weight(p,r,w_i)$$

$$\begin{aligned} SP(in, TMP, Texas) = {} & sim(Texas, November) \cdot weight(in, TMP, November) + \\ & sim(Texas, century) \cdot weight(in, TMP, century) + \\ & sim(Texas, month) \cdot weight(in, TMP, month) \end{aligned}$$

Selectional Preferences for SRL (Zapirain et al., 2013)

Distributional SP models: based on Erk's (2007) setting

JFK was assassinated [in Texas]$_{???}$

SP in–*TMP*: *November, century, month*

SP in–*LOC*: *Dallas, railway, city*

$$SP_{sim}(p, r, w) = \sum_{w_i \in Seen(p,r)} sim(w, w_i) \cdot weight(p, r, w_i)$$

$$\begin{aligned} SP(in, TMP, Texas) = \; & sim(Texas, November) \cdot weight(in, TMP, November) + \\ & sim(Texas, century) \cdot weight(in, TMP, century) + \\ & sim(Texas, month) \cdot weight(in, TMP, month) \end{aligned}$$

Selectional Preferences for SRL (Zapirain et al., 2013)

Distributional SP models: based on Erk's (2007) setting

JFK was assassinated [in Texas]$_{???}$

SP in–*TMP*: *November, century, month*

SP in–*LOC*: *Dallas, railway, city*

$$SP_{sim}(p,r,w) = \sum_{w_i \in Seen(p,r)} sim(w,w_i) \cdot weight(p,r,w_i)$$

$$\begin{aligned} SP(in, TMP, Texas) \; = \; & sim(Texas, November) \cdot freq(in, TMP, November) + \\ & sim(Texas, century) \cdot freq(in, TMP, century) + \\ & sim(Texas, month) \cdot freq(in, TMP, month) \end{aligned}$$

Selectional Preferences for SRL (Zapirain et al., 2013)

Distributional SP models: based on Erk's (2007) setting

JFK was assassinated [in Texas]$_{???}$

SP in–*TMP*: *November, century, month*

SP in–*LOC*: *Dallas, railway, city*

$$SP_{sim}(p, r, w) = \sum_{w_i \in Seen(p,r)} sim(w, w_i) \cdot weight(p, r, w_i)$$

$$\begin{aligned} SP(in, LOC, Texas) \;=\; & sim(Texas, Dallas) \cdot freq(in, LOC, Dallas) + \\ & sim(Texas, railway) \cdot freq(in, LOC, railway) + \\ & sim(Texas, city) \cdot freq(in, LOC, city) \end{aligned}$$

SP(in,LOC,Texas) > SP(in,TMP,Texas)

Selectional Preferences for SRL (Zapirain et al., 2013)

Distributional SP models: based on Erk's (2007) setting

JFK was assassinated [in Texas]$_{???}$

SP in–*TMP*: *November, century, month*

SP in–*LOC*: *Dallas, railway, city*

$$SP_{sim}(p,r,w) = \sum_{w_i \in Seen(p,r)} sim(w,w_i) \cdot weight(p,r,w_i)$$

$$\begin{aligned} SP(in, LOC, Texas) \; = \; & sim(Texas, Dallas) \cdot freq(in, LOC, Dallas) + \\ & sim(Texas, railway) \cdot freq(in, LOC, railway) + \\ & sim(Texas, city) \cdot freq(in, LOC, city) \end{aligned}$$

SP(in,LOC,Texas) $>$ SP(in,TMP,Texas)

Selectional Preferences for SRL (Zapirain et al., 2013)

Distributional SP models: various instantiations for *sim*

- Using Padó and Lapata's software (2007) for computing distributional similarity measures
  - ⇒ Run on the British National Corpus
  - ⇒ Optimal parameterization as described in the paper
  - ⇒ Jaccard, cosine and Lin's similarity measures: sim_{Jac}, sim_{cos} and sim_{Lin}
- Using the already available Lin's thesaurus (Lin, 1998)
  - ⇒ Direct and second order similarity: sim^{th}_{Lin}, sim^{th2}_{Jac} and sim^{th2}_{cos}
  - ⇒ Average of both directions similarity

Evaluation of SPs in isolation (Zapirain et al., 2013)

Setting: Assign role labels to argument head words based solely on SP scores

⇒ For each head word (w), select the role (r) of the predicate or preposition (p) which fits best the head word:
$R_{sim}(p, w) = \arg\max_{r \in Roles(p)} SP_{sim}(p, r, w)$

⇒ SPs based on (p, r, w) triples from CoNLL-2005 data

⇒ In-domain (WSJ) and out-of-domain (Brown) test sets CoNLL-2005

⇒ Lexical baseline model: for a test pair (p, w), assign the role under which the head (w) occurred most often in the training data given the predicate (p)

Evaluation of SPs in isolation

(Zapirain et al., 2013)

Setting: Assign role labels to argument head words based solely on SP scores

⇒ For each head word (w), select the role (r) of the predicate or preposition (p) which fits best the head word:
$R_{sim}(p, w) = \arg\max_{r \in Roles(p)} SP_{sim}(p, r, w)$

⇒ SPs based on (p, r, w) triples from CoNLL-2005 data

⇒ In-domain (WSJ) and out-of-domain (Brown) test sets CoNLL-2005

⇒ Lexical baseline model: for a test pair (p, w), assign the role under which the head (w) occurred most often in the training data given the predicate (p)

Evaluation of SPs in isolation (Zapirain et al., 2013)

Setting: Assign role labels to argument head words based solely on SP scores

⇒ For each head word (w), select the role (r) of the predicate or preposition (p) which fits best the head word:
$R_{sim}(p, w) = \arg\max_{r \in Roles(p)} SP_{sim}(p, r, w)$

⇒ SPs based on (p, r, w) triples from CoNLL-2005 data

⇒ In-domain (WSJ) and out-of-domain (Brown) test sets CoNLL-2005

⇒ Lexical baseline model: for a test pair (p, w), assign the role under which the head (w) occurred most often in the training data given the predicate (p)

Evaluation of SPs in isolation (Zapirain et al., 2013)

	WSJ-test			Brown		
	prec.	rec.	F_1	prec.	rec.	F_1
lexical	82.98	43.77	57.31	68.47	13.60	22.69
SP_{Res}	63.47	53.24	57.91	55.12	44.15	49.03
$SP_{sim_{Jac}}$	61.83	61.40	61.61	55.42	53.45	54.42
$SP_{sim_{cos}}$	64.67	64.22	64.44	56.56	54.54	55.53
$SP_{sim^{th2}_{Jac}}$	70.82	70.33	70.57	62.37	60.15	61.24
$SP_{sim^{th2}_{cos}}$	70.28	69.80	70.04	62.36	60.14	61.23

- ⇒ Lexical features have a high precision but very low recall
- ⇒ SPs are able to effectively generalize lexical features
- ⇒ SPs based on distributional similarity are better
- ⇒ Second-order similarity variants (Lin) attain the best results

Evaluation of SPs in isolation (Zapirain et al., 2013)

	WSJ-test			Brown		
	prec.	rec.	F_1	prec.	rec.	F_1
lexical	82.98	43.77	57.31	68.47	13.60	22.69
SP_{Res}	63.47	53.24	57.91	55.12	44.15	49.03
$SP_{sim_{Jac}}$	61.83	61.40	61.61	55.42	53.45	54.42
$SP_{sim_{cos}}$	64.67	64.22	64.44	56.56	54.54	55.53
$SP_{sim^{th2}_{Jac}}$	70.82	70.33	70.57	62.37	60.15	61.24
$SP_{sim^{th2}_{cos}}$	70.28	69.80	70.04	62.36	60.14	61.23

⇒ Lexical features have a high precision but very low recall

⇒ SPs are able to effectively generalize lexical features

⇒ SPs based on distributional similarity are better

⇒ Second-order similarity variants (Lin) attain the best results

SPs in a SRL System (Zapirain et al., 2013)

- *SwiRL* system for SRL (Surdeanu et al., 2007)
  - ⇒ System from CoNLL-2005 shared task (PropBank)
  - ⇒ Standard architecture (ML based on AdaBoost and SVMs)
  - ⇒ Best results from single (non-combined) systems at CoNLL-2005
- Simple approach: extending *SwiRL* features with SP predictions
  - ⇒ We train several extended *SwiRL-*SP_i models, one per selectional preferences model SP_i
  - ⇒ For each example (p, w) of *SwiRL-*SP_i, we add a single new feature whose value is the predicted role label $R_i(p, w)$

SPs in a SRL System (Zapirain et al., 2013)

Results

	WSJ-test			Brown		
	Core	Adj	All	Core	Adj	All
SwiRL	93.25	81.31	90.83	84.42	57.76	79.52
$SwiRL+SP_{Res}$	93.17	81.08	90.76	84.52	59.24	79.86
$SwiRL+SP_{sim_{Jac}}$	93.37	80.30	90.86	84.43	59.54	79.83
$SwiRL+SP_{sim_{cos}}$	93.33	80.92	90.87	85.14	60.16	80.50
$SwiRL+SP_{sim^{th2}_{Jac}}$	93.03	82.75	90.95	85.62	59.63	80.75
$SwiRL+SP_{sim^{th2}_{cos}}$	93.78	80.56	91.23	84.95	61.01	80.48

⇒ Slight improvements, especially noticeable on Brown corpus

⇒ Weak signal of a single feature?

SPs in a SRL System (Zapirain et al., 2013)

- Simple combinations of the individual $SwiRL{+}SP_i$ classifiers worked quite well (majority voting)
- We also trained a meta-classifier to combine the $SwiRL{+}SP_i$ classifiers and the stand-alone SP_i models:
  - ⇒ Binary classification approach: "is a proposed role correct or not?"
  - ⇒ Features are based on the predictions of base SP_i and $SwiRL{+}SP_i$ models
  - ⇒ Trained with a SVM with a quadratic polynomial kernel

SPs in a SRL System

(Zapirain et al., 2013)

Results (II)

	WSJ-test			Brown		
	Core	Adj	All	Core	Adj	All
SwiRL	93.25	81.31	90.83	84.42	57.76	79.52
$+SP_{sim_{cos}^{th2}}$	93.78	80.56	91.23	84.95	61.01	80.48
Meta	94.37	83.40	92.12	86.20	63.40	81.91

- Statistically significant improvements (99%) for both core and adjunct arguments, both in domain and out of domain

SPs in a SRL System

(Zapirain et al., 2013)

Results (II)

	WSJ-test			Brown		
	Core	Adj	All	Core	Adj	All
SwiRL	93.25	81.31	90.83	84.42	57.76	79.52
$+SP_{sim^{th2}_{cos}}$	93.78	80.56	91.23	84.95	61.01	80.48
Meta	94.37	83.40	92.12	86.20	63.40	81.91

- Statistically significant improvements (99%) for both core and adjunct arguments, both in domain and out of domain

SPs in a SRL System

(Zapirain et al., 2013)

Output analysis

- Manual inspection of 50 cases in which the meta classifier corrects SwiRL:
  - ⇒ Usually cases with low frequency verbs or argument heads
  - ⇒ In ~58% of the cases, syntax does not disambiguate, seems to suggest a wrong role label or it is confusing SwiRL because it is incorrect. However, most of the SP predictions are correct.
  - ⇒ ~30% of the cases: unclear source of the SwiRL error but still several SP models suggest the correct role
  - ⇒ ~12% of the cases: chance effect

SPs in a SRL System (Zapirain et al., 2013)

Output analysis

- Manual inspection of 50 cases in which the meta classifier corrects SwiRL:
  - ⇒ Usually cases with low frequency verbs or argument heads
  - ⇒ In ~58% of the cases, syntax does not disambiguate, seems to suggest a wrong role label or it is confusing SwiRL because it is incorrect. However, most of the SP predictions are correct.
  - ⇒ ~30% of the cases: unclear source of the SwiRL error but still several SP models suggest the correct role
  - ⇒ ~12% of the cases: chance effect

SPs in a SRL System (Zapirain et al., 2013)

Output analysis: example 1

		Several	JJ	(S1(S(NP*
		traders	NNS	*)
		could	MD	(VP*
		be	VB	(VP*
		seen	VBN	(VP*
		shaking	VBG	(S(VP*
		their	PRP$	(NP*
		heads	NNS	*)))
		when	WRB	(SBAR(WHADVP*)
A1	A0	the	DT	(S(NP*
A1	A0	news	NN	*)
	(P)	flashed	VBD	(VP*))))))
		.	.	*))

SPs in a SRL System (Zapirain et al., 2013)

Output analysis: example 2

		Italian	NNP	(S1(S(NP*
		President	NNP	*
		Francesco	NNP	*
		Cossiga	NNP	*)
	(P)	promised	VBD	(VP*
A2	A1	a	DT	(NP(NP*
A2	A1	quick	JJ	*
A2	A1	investigation	NN	*)
A2	A1	into	IN	(PP*
A2	A1	whether	IN	(SBAR*
A2	A1	Olivetti	NNP	(S(NP*)
A2	A1	broke	VBD	(VP*
A2	A1	Cocom	NNP	(NP*
A2	A1	rules	NNS	*)))))))
		.	.	*))

SPs in a SRL System

(Zapirain et al., 2013)

Output analysis: example 3

		Annual	JJ	(S(NP*
		payments	NNS	*)
		will	MD	(VP*
		more	RBR	(VP(ADVP*
		than	IN	*)
	(P)	double	VB	*
A3	TMP	from	IN	(PP*
A3	TMP	a	DT	(NP*
A3	TMP	year	NN	*
A3	TMP	ago	RB	*))
		to	TO	(PP*
		about	RB	(NP(QP*
		$240	CD	*
		million	CD	*)))
		...		

SPs in a SRL System (Zapirain et al., 2013)

Output analysis: example 4

		Procter	NNP	(S1(S(NP*
		&	CC	*
		Gamble	NNP	*
		Co.	NNP	*)
		plans	VBZ	(VP*
		to	TO	(S(VP*
		begin	VB	(VP*
	(P)	testing	VBG	(S(VP*
		next	JJ	(NP*
		month	NN	*)))
A1	A0	a	DT	(NP(NP*
A1	A0	superco.	JJ	*
A1	A0	detergent	NN	*)
A1	A0	that	WDT	(SBAR(WHNP*)
		...		
A1	A0	washload	NN	(NP*))))))))))))
		.	.	*))

Talk Overview

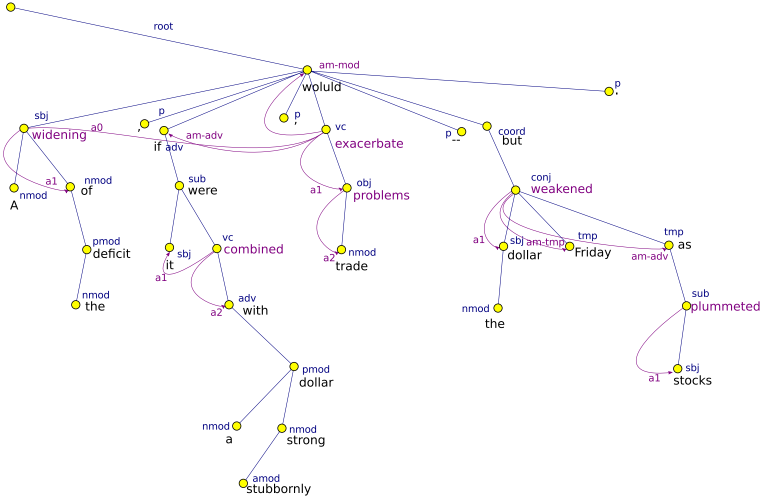
Joint work with

Xavier Lluís and *Xavier Carreras*

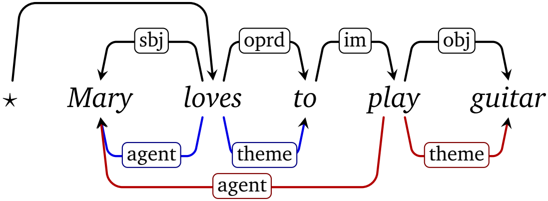
(Lluís et al. 2013) — TACL (to be presented at ACL)

CoNLL-2008/2009 shared task

Joint parsing of syntactic and semantic dependencies

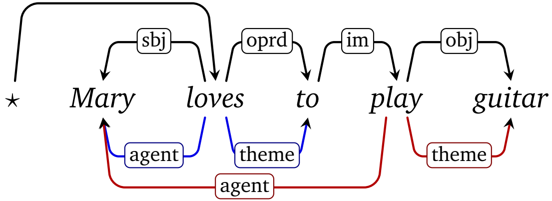


A Simplified Example



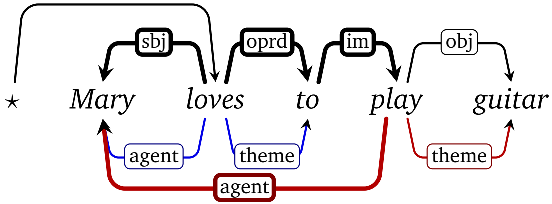
- Predicate-argument structures are naturally represented with dependencies

A Simplified Example



- Semantic roles are strongly related to syntactic structure
- Typical systems find semantic roles in a pipeline
  - ⇒ First obtain the syntactic tree
  - ⇒ Second obtain the semantic roles, using the syntactic tree
- Pipeline systems can not correct syntax based on semantic roles

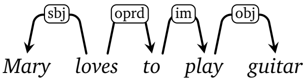
A Simplified Example



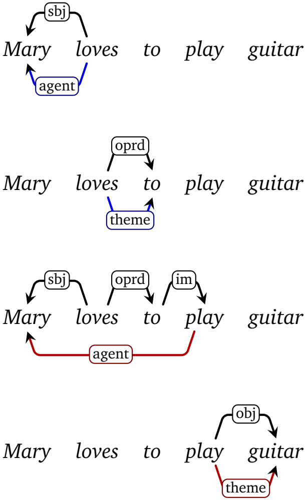
- We model the two structures jointly
  - ⇒ To capture interactions between syntactic and semantic dependencies
- Challenge:
  - ⇒ Some semantic dependencies are associated with a segment of syntactic dependencies
  - ⇒ Hard to factorize the two structures jointly

Decomposing Syntactic and Semantic Trees

Syntactic Tree



Semantic Trees with Local Syntax



Semantic trees need to agree with the syntactic tree.

Semantic features can conjoin

- any syntactic feature with
- a semantic role

Syntactic subproblem

$$\begin{aligned} syn(\mathbf{x}) &= \operatorname*{argmax}_{\mathbf{y}} \ \text{score\_syn}(\mathbf{x}, \mathbf{y}) \\ \text{subject to} & \quad \text{cTree: } \mathbf{y} \text{ is a projective tree} \end{aligned}$$

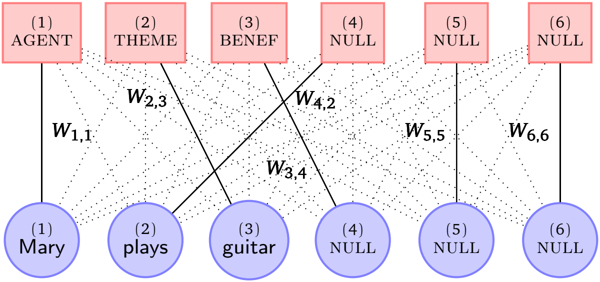
- Solved by a standard dependency parsing algorithm
- score_syn$(\mathbf{x}, \mathbf{y})$ is arc-factored: 1st and 2nd order models
- Graph-based parsing algorithms, reimplementing (McDonald, 2005; Carreras et al., 2007)
- Trained with (linear) average structure perceptron using state-of-the-art features

Semantic Subproblem

$$
\begin{aligned}
srl(\mathbf{x}) \quad &= \quad \operatorname*{argmax}_{\mathbf{z},\boldsymbol{\pi}} \ \text{score\_srl}(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}) \\
\text{subject to} \quad & \text{cRole: no repeated roles} \\
& \text{cArg: at most one role per token} \\
& \text{cPath: } \boldsymbol{\pi} \text{ codifies paths consistent with } \mathbf{z}
\end{aligned}
$$

- In a predicate:
  - ⇒ A token appears at most once as argument
  - ⇒ A semantic role appears at most once
- score_srl$(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi})$ is factorized at the level of $\langle \mathbf{x}, p, a, r, \pi^{p,a,r} \rangle$
- local score_srl$(\mathbf{x}, p, a, r, \pi^{p,a,r})$ provided by linear classifiers
- We frame the *argmax* inference as a *linear assignment problem*

SRL as Assignment



- The Hungarian algorithm solves it in $\mathcal{O}(n^3)$
- $w_{i,j}$ are the previous local predictions score_srl$(\mathbf{x}, p, a, r, \pi^{p,a,r})$
- In practice, the list of most likely paths from p to a is pre-computed using syntactic models
- Learning is performed with structure perceptron, with feedback applied after solving the assignment problem

Joint Syntactic-Semantic Inference

$$\begin{aligned} \langle \mathbf{y}^*, \mathbf{z}^*, \boldsymbol{\pi}^* \rangle \quad &= \quad \underset{\mathbf{y},\mathbf{z},\boldsymbol{\pi}}{\operatorname{argmax}} \ \mathrm{sc\_syn}(\mathbf{x}, \mathbf{y}) + \mathrm{sc\_srl}(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}) \\ \text{subject to} \quad & \text{cTree, cRole, cArg, cPath} \\ & \text{cSubtree: } \mathbf{y} \text{ is consistent with } \boldsymbol{\pi} \end{aligned}$$

Joint Syntactic-Semantic Inference

$$\begin{aligned}\langle \mathbf{y}^*, \mathbf{z}^*, \boldsymbol{\pi}^* \rangle \quad = & \quad \underset{\mathbf{y},\mathbf{z},\boldsymbol{\pi}}{\operatorname{argmax}} \; \mathrm{sc\_syn}(\mathbf{x}, \mathbf{y}) + \mathrm{sc\_srl}(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}) \\ \text{subject to} & \quad \text{cTree}, \; \text{cRole}, \; \text{cArg}, \; \text{cPath} \\ & \quad \text{cSubtree: } \mathbf{y} \text{ is consistent with } \boldsymbol{\pi}\end{aligned}$$

cSubtree constraints can be easily expressed as:

$$\forall d \in \mathbf{y}, \quad c \cdot \mathbf{y}_d \geqslant \sum_{p,a,r \in \mathbf{z}} \boldsymbol{\pi}_d^{p,a,r}$$

or, equivalently, as equality constraints

$$\forall d \in \mathbf{y}, \quad c \cdot \mathbf{y}_d - \sum_{p,a,r \in \mathbf{z}} \boldsymbol{\pi}_d^{p,a,r} - \xi_d = 0$$

Joint Syntactic-Semantic Inference

- We employed Dual Decomposition to solve the joint inference (Rush and Collins, 2011) (Sontag et al 2010)
- Lagrangian relaxation-based method that iteratively solves decomposed sub-problems with agreement constraints:
  - ⇒ Subtree constraints are relaxed by introducing Lagrange multipliers for every dependency λ_d
  - ⇒ Subproblems now depend on the $\boldsymbol{\lambda}$ penalty variables but can be efficiently solved
  - ⇒ Syntax: standard dependency parsing inference
  - ⇒ Semantic: linear assignment
- Guaranteed optimal solution when it converges
- In experiments, convergence in $> 99.5\%$ of sentences

Experiments and Results

We ran experiments on the CoNLL-2009 datasets with the following configurations:

- **Pipeline** best *syn* then best *srl* enforcing cArg
- **+Assignment** enforces cRole, cArg over best *syn*
- **Forest** works with a forest of *syn* trees
- **DD** applies dual-decomposition

Experiments and Results

	syn	sem		
system	acc	prec	rec	F_1
Pipeline-1				
+Assignment-1				
Forest-1				
DD-1				

Results on WSJ development set

Experiments and Results

	syn	sem		
system	acc	prec	rec	F_1
Pipeline-1	85.32	86.23	67.67	75.83
+Assignment-1	85.32	84.08	71.82	77.47
Forest-1				
DD-1				

+Assignment improves over *Pipeline*

Experiments and Results

	syn	sem		
system	acc	prec	rec	F_1
Pipeline-1	85.32	86.23	67.67	75.83
+Assignment-1	85.32	84.08	71.82	77.47
Forest-1	85.32	80.67	73.60	76.97
DD-1				

Forests shows higher recall

Experiments and Results

	syn	sem		
system	acc	prec	rec	F_1
Pipeline-1	85.32	86.23	67.67	75.83
+Assignment-1	85.32	84.08	71.82	77.47
Forest-1	85.32	80.67	73.60	76.97
DD-1	85.48	83.99	72.69	77.94

DD-1 achieves better sem F_1

Experiments and Results

	syn	sem		
system	acc	prec	rec	F_1
Pipeline-1	85.32	86.23	67.67	75.83
+Assignment-1	85.32	84.08	71.82	77.47
Forest-1	85.32	80.67	73.60	76.97
DD-1	85.48	83.99	72.69	77.94
Pipeline-2	87.77	87.07	68.65	76.77
+Assignment-2	87.77	85.21	73.41	78.87
Forest-2	87.77	80.67	73.60	76.97
DD-2	87.84	85.20	73.23	78.79

Second-order paths are quite accurate

Experiments and Results

	syn	sem			
WSJ	acc	prec	rec	F_1	PP
Lluís09	87.48	73.87	67.40	70.49	39.68
Merlo09	88.79	81.00	76.45	78.66	54.80
DD-2	89.21	86.01	74.84	80.04	55.73

Results in WSJ corpus (in-domain) test set

Experiments and Results

	syn	sem			
WSJ	acc	prec	rec	F_1	PP
Lluís09	87.48	73.87	67.40	70.49	39.68
Merlo09	88.79	81.00	76.45	78.66	54.80
DD-2	89.21	86.01	74.84	80.04	55.73

Better results than *Merlo09*

Experiments and Results

	syn	sem			
Brown	acc	prec	rec	F_1	PP
Lluís09	80.92	62.29	59.22	60.71	29.79
Merlo09	80.84	68.97	63.06	65.89	38.92
DD-2	82.61	74.12	61.59	67.83	38.92

Results in Brown corpus (out-of-domain) test set

Thank you!

Exploring challenges in Semantic Role Labeling

Lluís Màrquez
TALP Research Center
Tecnhical University of Catalonia

Invited talk at ABBYY Open Seminar
Moscow, Russia, May 28, 2013