Enhancing and Applying Dependency Analysis in Different Topics: Negation, Speculation and Text Simplification

Miguel Ballesteros

Universidad Complutense de Madrid

miballes@fdi.ucm.es

October 20, 2011

Outline

- Who am I?
- My background.
- My goals.
- My own state of the art.
- My PhD thesis project.

Who am I?

1. I am a Doctoral Candidate in Department of Software Engineering and Artificial Intelligence at Universidad Complutense de Madrid, where I am developing my thesis about this very thing: "Analyzing, Enhancing and Applying Dependency Parsing".
2. I hold a B.S from the Universidad Complutense de Madrid and a MsC at the same University.

My Degree (2009)

1. I studied Computer Science (or Computer Engineering) in Madrid. Which is a 5 years degree (it is like degree + master).
2. Therefore, my background is completely computational.
3. My final degree project was a tool for Authomata Theory, Regular Expressions and Formal Languages.

My Master Thesis (2010)

1. After my 5 years degree, I studied a Research in Informatics Master (which is the way to start a PhD thesis–old PhD courses).
2. I had Computational Linguistics, Case Based Reasoning, Machine Learning and several different theoretical subjects.
3. My final master thesis was a study on Dependency Parsing using MaltParser for Spanish, studying the consistency of the accuracy and also some steps trying to enhance the accuracy.

- My Advisors
  - Pablo Gervás.
  - Virginia Francisco.
- People who are helping me.
  - Jesús Herrera.
  - Alberto Díaz.
  - Joakim Nivre.

My PhD thesis (2010-...)

My Topic(s): Dependency Parsing.

1. Analyzing Dependency Parsing.
2. Enhancing Dependency Parsing.
3. Applying Dependency Parsing.

Analyzing, Enhancing and Applying Dependency Parsing.

Applying Dependency Parsing
Outline

So far, I have applied Dependency syntactic structures to "solve" the following.

1. Terms affected by negation signals.
2. The Scope of Negation.
3. The Scope of Speculation.
4. Simplify Sentences using a pruning tree algorithm.

Terms Affected by Negation Signals

Computer Cooking Contest Project

1. I was involved in a Computer Cooking Contest project.
2. The idea was to develop a system that suggest recipes using a given list of ingredients.
3. Our input was Natural Language Sentences. Since my master thesis was about dependency parsing, we tried to infer the terms affected by negation signals using dependency structures.

Terms Affected by Negation Signals

Computer Cooking Contest Project

1. We used the Minipar parser (Dekang Lin).
2. At this time, I had no experience with Machine learning parsers.
3. It worked ok, therefore, we tried to go on with it.

I want to eat rice, saffron, shrimps, chicken, crab, squid but I hate apples.

With this query the system returned the following recipes:

- *Spanish Paella*
- *Seafood Bouillabaisse*
- *Brown Rice Jambalaya*
- ...

1. This work was published in the Computer Cooking Contest at the International Conference on Case Based Reasoning.
2. We obtained the best student paper award.
3. The system that we presented there can be accesed via `http://minerva.fdi.ucm.es:8888/CCC2010/`.

- We realized that the very same algorithm used in the CCC can be used to Infer the Scope of (neg and spec) Signals.
- Every text contains information that includes uncertainty, deniability or speculation.
- It is important to distinguish between speculative/negative statements and factual ones.
- Chapman et al. (2002) proved that in a search for *fracture* in a radiology reports database, 95 to 99 percent of the reports returned would state *"no signs of fracture"* or words to that effect.

Inferring the Scope of Negation and Speculation

- We build a system in which the domain application is somewhat open using a different lexicon of cues.
- **Affected Wordforms Detection Algorithm**: an algorithm that detects wordforms within the scope of cues based on dependency Parsing. (COOKING CONTEST!)
- **Scope Finding algorithm**: it uses the output of the Affected Wordforms Detection Algorithm to annotate sentences with the scope of cues. (POST–PROCESSING)

Inferring the Scope of Negation and Speculation

Bioscope Corpus

- It is a standard annotated with the scope of negation and speculation.
- Divided in biomedical scientific papers, abstracts and clinical reports.

```
<sentence id="S1.123">
        The reason why
        <xcope id="X1.123.1">
                the two other families were
                <cue type="negation" ref="X1.123.1">
                        not
                </cue>
                detected
        </xcope>
        is more complex.
</sentence>
```

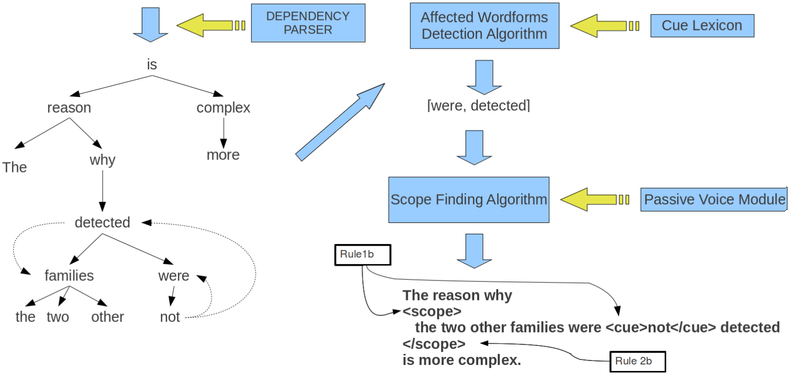
Figure: A sentence annotated with the scope of negation in the Bioscope corpus.

Negation and Speculation Cue Lexicon

not	no	neither..nor	none
discard	rule out	fail	avoid
absence	lack (v)	lack (n)	without
unable	rather than	absent	can not

appear	can	could	either
indicate that	indicate	imply	evaluate for
likely	may	might	or
possible	possibly	potential	potentially
propose	putative	rule out	suggest
think	unknown	whether	would

The reason why the two other families were not detected is more complex.



Collection	System	Precision	Recall	F1	PCS	PCNC
Papers	Our Results	73.49%	**80.70%**	**76.93%**	**56.43%**	91.15%
	Morante et al.	72.21%	69.72%	70.94%	41.00%	**92.15%**
	Zhu et al.	56.27%	58.20%	57.22%	–	–
	Councill et al.	**80.80%**	70.80%	75.50%	53.70%	–
Abstracts	Our Results	**84.92%**	**84.03%**	**84.48%**	**68.92%**	**95.56%**
	Morante et al.	81.76%	83.45%	82.60%	66.07%	95.09%
	Zhu et al.	78.24%	78.77%	78.50%	–	–
Clinical	Our Results	**95.83%**	**90.58%**	**93.13%**	**89.06%**	94.82%
	Morante et al.	86.38%	82.14%	84.20%	70.75%	**97.72%**
	Zhu et al.	82.22%	80.62%	81.41%	–	–

Speculation Results

Collection	System	Precision	Recall	F1	PCS	PCHC
Papers	Our Results	**82.78%**	**73.88%**	**78.08%**	**39.43%**	80.38%
	Morante et al.	67.97%	53.16%	59.66%	35.92%	**92.15%**
	Zhu et al.	56.27%	58.20%	57.22%	–	–
Abstracts	Our Results	**87.96%**	**75.35%**	**81.14%**	46.75%	79.50%
	Morante et al.	85.77%	72.44%	78.54%	**65.55%**	**96.03%**
	Zhu et al.	81.58%	73.34%	77.24%	–	–
Clinical	Our Results	**83.96%**	**67.15%**	**74.62%**	**36.20%**	**67.19%**
	Morante et al.	68.21%	26.49%	38.16%	26.21%	64.44%
	Zhu et al.	70.46%	25.59%	37.55%	–	–

- These systems are also online and can be accessed via:
  - `http://minerva.fdi.ucm.es:8888/ScopeTagger`
  - `http://minerva.fdi.ucm.es:8888/ScopeTaggerSpec`
- It is also published in KDIR 2011 (29th october! *Speculation*), SEPLN 2011 (Demo session) and I have a submission waiting for revision (*Negation*).

- An accurate negation and speculation scope classification system is really useful.
- Our positive results show that dependency Parsing is useful for detecting negation and speculation.
- The sentences involved in Bioscope are relatively complex with very different syntactic structures, but our system is able to accurately detect negations and speculations and their scopes inside them.
- The domain is open, we demonstrated it changing the task and the domain twice.

Main goal

Our goal was to build a system to promote access to Spanish texts for people at the rudimentary and basic literacy levels, as well as for those with cognitive disabilities.

- Can we use dependency parsing to do that?

Text Simplification
Motivation

- Long sentences, conjoined sentences, embedded clauses, passives, non-canonical word order and use of low-frequency words increase text complexity.
- We focused on the syntactic structure of a text to maximize the comprehension of written texts through the simplification of their linguistic structure.
- There are guidelines to make text easier to read and comprehend.

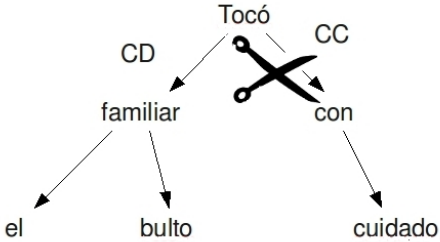
- 3500 sentences corpus.
- It is used as normal corpus for Spanish Dependency Parsing.
- CoNLL data format.
- Sentences from newspapers, literary sentences, etc.

- We propose a rule-based syntactic simplification system.
- It uses as input a dependency parsed tree.
- Using the output of a multilingual dependency parser, like Maltparser, you can simplify any sentence in Spanish.
- This system is in a very very first step.

- We were wondering which tag is the most appropriate to be removed.
- There is a small subset of tags that can be removed without losing the main information of the sentence.
- We decided to remove complementary information about an action, like when, where, how and why.
- But we are not always losing this kind of information.

Text Simplification

Dependency Tree Pruning



Text Simplification

Example: Input

Tocó el familiar bulto con cuidado, recorriendo sus aristas con las yemas de los dedos, contemplando la imagen que le devolvía el espejo y pensando que todo aquello ya no tenía remedio , que nada podía hacer ya por su cara , ni por su pecho , por esas piernas que no veía , pero sabía tan huesudas y separadas como las patas de un pollo mojado , y por esa carne blanquecina , fofa , que comenzaba a acumularse en torno a su cintura , a descolgarse hacia abajo arrastrando en su vértigo un ombligo progresivamente hondo , para añadir una nueva vejación , la de los años , a un cuerpo condenado de antemano , desde antes de existir , a ser feo.

Text Simplification

Example: Output

Tocó el familiar bulto , recorriendo sus aristas , contemplando la imagen que le devolvía el espejo y pensando que todo aquello no tenía remedio, que nada podía hacer.

- Our system removes a lot of extra information for this sentence.
- The simplified version keeps the main information and it is grammatically correct.
- The simplified version is easier to read than the original version.

- Two measures of evaluation:
  1. Questionnaire for adults .
  2. Overall statistics in corpus.

We surveyed a group of people (20) about how good was the text simplification made.

- We asked them 4 questions about the sentences.
- None of them know how the simplification algorithm works.
- We showed them the whole sentence and the simplified version.

- Q1: *Is the main idea of the sentence retained?*
- Q2: *Was all the removed information unnecessary?*
- Q3: *Have only details without importance been deleted?*
- Q4: *Do you understand better the simplified sentence than the normal sentence?*

Question	YES	NO
Q1	67.58%	32.42%
Q2	27.66%	72.34%
Q3	46.72%	53.28%
Q4	60.76%	39.24%

Table: Results obtained by the survey.

- We simplified the whole corpus to find a global average of simplification.
- The algorithm simplified 2,737 sentences of 3,512 sentences because some of them are already simplified in the original corpus.

	Original	Simplified
Total Wordforms	95,028	58,415
Average SL	27.06 wf	16.63 wf
Longest SL	143 wf	94 wf

Table: Results on Sentence Length (SL)

- The potentialities of text simplification systems for education are obvious.
- The social impact of text simplification is undeniable.
- Our system is a first approximation.
- It is possible to simplify sentences using dependency parsing.

So far, I have tried the following studyng dependency parsing towards an enhancement of the accuracy.

1. Studies about the training corpora.
2. Enhancing of Accuracy combining small trained specific parsers.
3. Whole Parsing Combination.

Studies about the Training Corpora

- Manipulation of the training corpora to find if the accuracy is homogeneus.
- We found that the accuracy is homogeneus, but we detected some important things.
- We must build the training corpora carefully because there is extra information.
- We realized that complete–match accuracy is not very high. For some purposes it is very important.

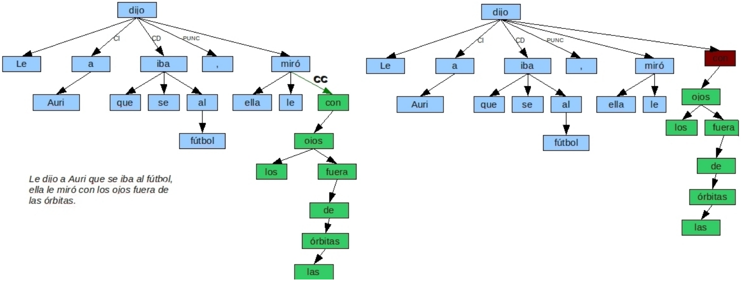
Enhancing of Accuracy combining small trained specific parsers.

The Case of Spanish

- In order to improve the accuracy we did the following:
- There is a small set of words that are more frequently incorrectly parsed:
  - The conjunction (y/e).
  - The prepositions 'a', 'de', 'en', 'con', 'por'.
  - The nexus 'que'.
- These words produce most of the errors.
- Can we reduce this percentage?

Why these words are important?

They are function words and as we can see here, they are really important.



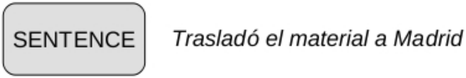
Enhancing of Accuracy combining small trained specific parsers.

Our Proposal

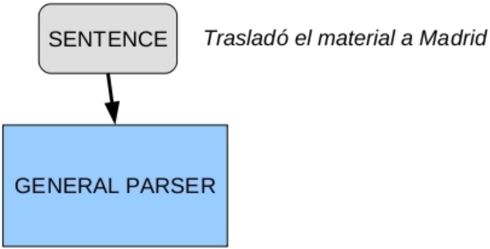
Automatic generation of N specific parsers trained to parse these words, combining the action of them with a general parser trained with the whole training corpus.

- N different parsers for each word (M words), finally we have (N x M) + 1 different parsers.
- Each one trained in a different way, with an specific feature model.
- Each one trained with an specific automatic built corpus from the whole corpus.

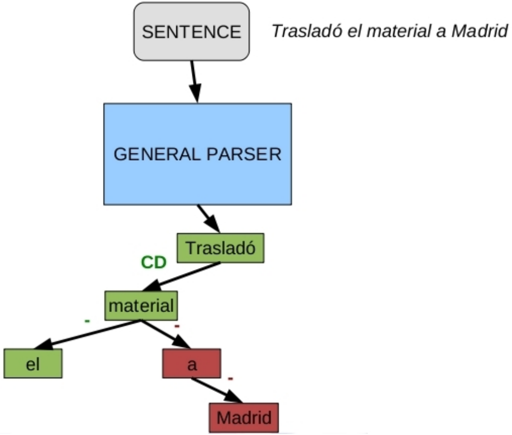
Combining small trained specific parsers.



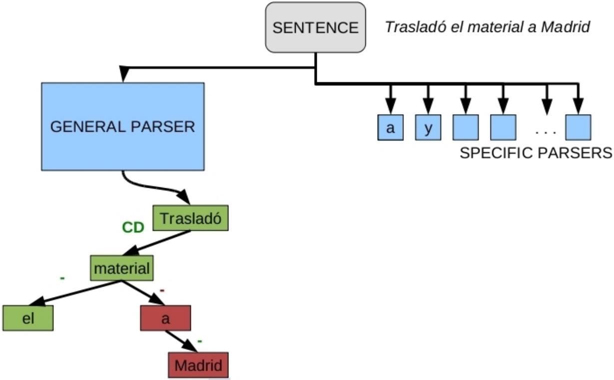
Combining small trained specific parsers.



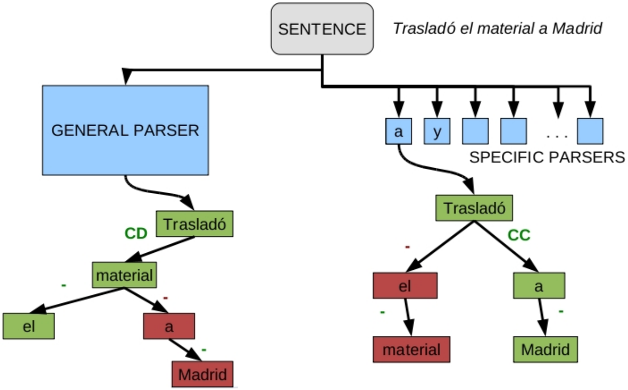
Combining small trained specific parsers.



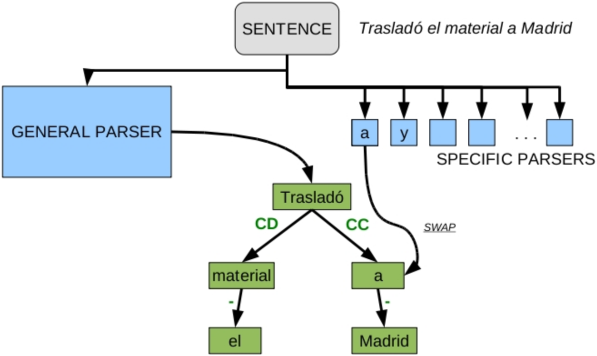
Combining small trained specific parsers.



Combining small trained specific parsers.



Combining small trained specific parsers.



Combining small trained specific parsers.
Conjunction results

Case	#1	#2	#3	#4
Label	–	–	–	–
Connected with	verb$^{\leftarrow}$	propper noun$^{\leftarrow}$	common noun$^{\leftarrow}$	adjective$^{\leftarrow}$
LAS$_{y/e}$ before	**81.3%**	80%	66.7%	80%
LAS$_{y/e}$ after	75%	**100%**	**80%**	**100%**
# Sentences train corpus	361	59	266	59

Combining small trained specific parsers.

Preposition 'a' results

Case	#1	#2	#3	#4	#5	#6
Label	CD	CI	CC	CREG	–	–
Connected with	verb$^{\leftarrow}$					noun$^{\leftarrow}$
LAS$_a$ before	62.5%	42.9%	60%	25%	0%	50%
LAS$_a$ after	**87.5**%	**100**%	**100**%	**75**%	0%	**100**%
# Senteces train corpus	110	80	146	86	8	63

Combining small trained specific parsers.

Preposition 'de' results

Case	#1	#2	#3	#4
Label	CC	CREG	–	–
Connected with	verb$^{\leftarrow}$		adverb or adjective$^{\leftarrow}$	noun$^{\leftarrow}$
LAS$_{de}$ before	0%	0%	**100%**	83.3%
LAS$_{de}$ after	**100%**	**100%**	**100%**	**96.7%**
# Sentences train corpus	535	105	39	32

Combining small trained specific parsers.
Nexus 'que' results

Case	#1	#2	#3
Label	SUJ	–	SUJ
Connected with	$\text{verb}^{\rightarrow}$		$\text{verb}^{\leftarrow}$
LAS$_{que}$ before	82.5%	86.4%	0%
LAS$_{que}$ after	**92.3%**	**95.5%**	**100%**
# Sentences train corpus	349	342	6

Combining small trained specific parsers.

Preposition 'en' results

Case	#1	#2	#3	#4
Label	CC	CC	CREG	–
Connected with	verb$^{\rightarrow}$	verb$^{\leftarrow}$		noun$^{\leftarrow}$
LAS$_{en}$ before	**83.3%**	92.6%	50%	62.5%
LAS$_{en}$ after	**83.3%**	**100%**	**100%**	**87.7%**
# Sentences train corpus	111	363	55	121

Combining small trained specific parsers.

Preposition 'con' results

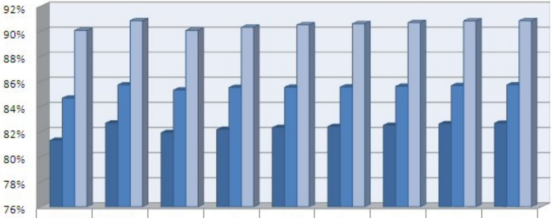
Case	#1	#2	#3	#4
Label	CC	CREG	–	–
Connected with	verb$^{\leftarrow}$			noun$^{\leftarrow}$
LAS$_{con}$ before	60%	40%	**100%**	66.7%
LAS$_{con}$ after	**80%**	**100%**	**100%**	**83.3%**
# Sentences train corpus	204	39	5	95

Combining small trained specific parsers.

Preposition 'por' results

Case	#1	#2	#3
Label	–	CAG	CAG
Connected with	noun$^{\leftarrow}$	comma$^{\leftarrow}$	adjective$^{\leftarrow}$
LAS$_{por}$ before	**100%**	**100%**	80%
LAS$_{por}$ after	**100%**	**100%**	**100%**
# Sentences train corpus	47	13	71

- 28 specific parsers.
- We obtain better results in 27 of the 28 cases.



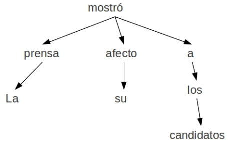
	Baseline	Overall	y/e	a	de	que	en	con	por
LAS	81,30%	82,67%	81,92%	82,17%	82,32%	82,40%	82,50%	82,62%	82,67%
UAS	84,67%	85,73%	85,31%	85,51%	85,53%	85,55%	85,61%	85,67%	85,73%
LA	90,06%	90,84%	90,06%	90,32%	90,50%	90,60%	90,70%	90,82%	90,84%

- Our proposal is feasible.
- The resultant trees are better built.
- The local accuracy is much better.
- Is it possible to build an automatic algorithm?

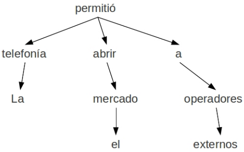
- We have built an algorithm that follows the combination process automatically.
- The obtained algorithm, is capable to send the correct sentences to the best parser, but not always.
- Corpus inconsistencies.
- Bad approach! :(

Combining small trained specific parsers.

Problems found in the corpus



La prensa mostró su afecto a los candidatos [The press showed its affection towards the candidates]

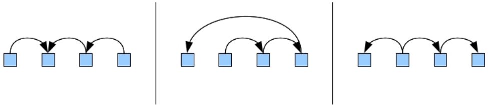


La telefonía permitió abrir el mercado a operadores externos [The telephone system opened the market to external operators]

- This kind of combination is feasible.
- It is an easy and novel proposal.
- The corpus should be rebuilt carefully.
- The results are not very high.
- It was published in TSD 2010.

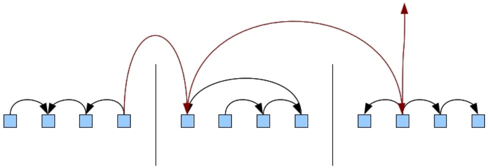
Finally, this is what I am trying to do right now.

1. Sentence segmentation in order to avoid error parsing propagation. (Uppsala!)
2. Whole Parsing Combination. (?)
3. Trying to fix my initial parsing combination. (?)



Two main questions

1. **We need to decide which positions are the best to split the sentences.**
2.



Two main questions

1. We need to decide which positions are the best to split the sentences.
2. **How can we get the long distance dependencies?**

Conclusions and Future Work

- The application branch of my thesis is closed. At least for me, but
  - I am involved in the organization of future Surface Realisation challenges using Dependency Parsing.
  - I am a "counselor" in a final degree project which is been advised by Alberto Díaz, who is a Professor in my University (about Speculation Scope classification).
  - I am not closing the door, but...
- Nowadays, I am more interested in the Enhancing of Statistical parsers.
  - I would like to publish the strong papers of my thesis in this topic.
  - A future system developed by myself?
  - This is why I am here! :-)

THANKS

1. http://nil.fdi.ucm.es
2. http://nil.fdi.ucm.es/index.php?q=node/449
3. miballes@fdi.ucm.es