Question Parsing for QA in Spanish

Iria Gayo
University of Santiago de Compostela
iria.delrio@usc.es

Abstract

Question processing is a key step in Question Answering systems. For this task, it has been shown that a good syntactic analysis of questions helps to improve the results. However, general parsers seem to present some disadvantages in question analysis. We present a specific tool under development for Spanish question analysis in a QA context: SpQA. SpQA is a parser designed to deal with the special syntactic features of Spanish questions and to cover some needs of question analysis in QA systems such as target identification. The system has been evaluated together with three Spanish general parsers. In this comparative evaluation, SpQA shows the best results in Spanish question analysis.

1 Introduction

In Question Answering (QA) systems, question processing is a crucial step to obtain a right answer (Carvalho et al., 2010). For this reason, QA systems usually have a specific module that addresses question analysis (Vicedo, 2004). Question treatment can have different levels of complexity, but, in most cases, it entails a syntactic analysis. Furthermore, in this analysis, the correct processing of the interrogative constituent has a special relevance, taking into account that this element can play an important role in the definition of the question target. Correct syntactic analysis of questions constitutes, therefore, a key stage in QA systems process (Moldovan et al., 2002; Hermjakob, 2001): if we want a good processing of the question, a good syntactic analysis is a helpful starting point.

Consequently, in order to get a good syntactic analysis of questions, we need a tool for processing them correctly. For Spanish there are free general parsers that could carry out this task. However, this option presents some disadvantages as we will see in detail in section 2. In this paper we present a tool under development for Spanish question analysis in a QA context, SpQA (Spanish Parser for QA). SpQA is a parser focused on question analysis, designed to be used in the question processing module of a QA system. As a result, it is thought to deal with the special syntactic features of Spanish questions and to cover some needs of QA systems such as question target identification. The parser has been evaluated comparing its results with those presented for general Spanish parsers in Gayo (2011). As we will see in section 5, this comparative evaluation shows that, currently, SpQA gets the best results in syntactic analysis of questions.

The paper is structured as follows: in section 2 we show some data related to the performance of general parsers in question analysis. In section 3 we briefly present SpQA. Section 4 accounts for the evaluation method and section 5 shows the results of this evaluation. Finally, in section 6 we present some conclusions and future work.

2 Parsing Questions

To confront the task of question analysis, we could think to make use of available general parsers. However, this option carries some drawbacks.

It has been shown, at least for English, that parsing accuracies of general parsers drop significantly on out-of-domain data (Gildea, 2001; Mcclosky et al., 2006; Foster, 2010). This fact has also been shown, in particular, for question analysis in English (Petrov et al., 2010).

Unfortunately, for Spanish there are not such studies that compare general accuracies of available parsers with those obtained parsing questions. Therefore, in order to obtain this kind of data, we can use available studies that measure question parsing performance and comparing them with others that measure general performance.

Related with question parsing in Spanish, we have only the data of Gayo (2011). Gayo (2011)

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 73–78,
Hissar, Bulgaria, 13 September 2011.

shows the accuracy in question analysis of three general Spanish parsers: DepPattern (Gamallo and Sánchez, 2009), Txala (Atserias et al., 2005) and Hispal (Bick, 2006). As we will see, this evaluation uses PARSEVAL metrics for two variables: constituent recognition and constituent labeling. These variables are applied to all consituents in the question and to the interrogative constituent in particular.

We can see the results of Gayo (2011) summarized in Table 1.

	Hispal	**Txala**	**DepPatt.**
All-Recognition	87.8	91.6	86.1
Int-Recognition	97.0	100.0	90.0
All-Labeling	68.2	71.3	51.1
Int-Labeling	52.5	62.0	25.0

Table 1: Evaluation of three Spanish parsers in question analysis Gayo (2011).

On the other hand, there are general evaluations only for two of the three parsers measured in Gayo (2011): Hispal (Bick, 2006) and Txala (Lloberes et al., 2010). Comparing these general results with those for questions showed in Table 1, we obtain the following data:[1]

	Txala	**Hispal**
G-Recognition	81.1/80.9	
Q-Recognition	91.6	87.8
Qint-Recognition	100.0	97.0
G-Labeling	73.9/74.3	95.3
Q-Labeling	71.3	68.2
Qint-Labeling	62.0	52.5

Table 2: Comparison of results in general (G) and question analysis (Q for all constituents; Qint for the interrogative contituent) of two Spanish parsers.

Because we do not have data for Hispal about general constituent recognition (see note 2), it is only possible to make an exhaustive comparison of both parsers concerning labeling. As we can see in Table 2, compared with general labeling (G-Labeling), accuracies of both parsers drop in tasks of question labelling (Q-Recognition and Qint-Recognition). This decrease is especially marked labeling the interrogative constituent (Qint-Labeling). However, the distance between accuracies in general and question analysis is considerably bigger in Hispal than in Txala. In fact, Txala only shows a remarkable drop labeling the interrogative constituent. We can conclude that Hispal seems to suffer considerably the change of domain, whereas Txala only shows some problems when it is confronted with one specific aspect of questions syntax: the role of the interrogative constituent.

3 SpQA

SpQA is a parser (under development) designed for Spanish question analysis in a QA context. Therefore, it is thought to be part of the question analysis module of QA systems. SpQA is a rule-based parser/transducer, wich is generated by means of the AGFL parser generator from an attribute grammar written in the AGFL formalism[2]. This grammar (with its lexicons and fact tables) is an extension of a general Spanish grammar for IR applications that is also under development, ASPIRA. The generated parser is a Top-Down Chart parser, using the Best-Only heuristic (Koster et al., 2007). It can perform constituent and dependency analysis (the latter by transduction).

The aim of the parser is to obtain as much linguistic information as possible from questions to facilitate the extraction of the right answer in a QA system. For this reason, we are interested in syntactic as well as semantic information, although, for the time being, the parser gets mostly syntactic information. Concerning the type of questions to analyze, we want to cover all types of Spanish direct interrogative structures (wh- and yes/no questions). At the current stage of development, the parser analyzes only wh- questions like:

¿Qué dibujó Leonardo Da Vinci en 1492?

(What did Leonardo Da Vinci draw in 1492?)

Given a question like this, SpQA

- recognizes and labels all the syntactic constituents in the sentence, showing the dependency relations between constituents
- identifies the syntactic and semantic target of the question (qt)

[1] In Lloberes et al. (2010), Txala was evaluated with two different corpora, so there are two different results. Unfortunately, Bick (2006) does not show general results for constituent identification (G-Recognition).

[2] http://www.agfl.cs.ru.nl/

- recognizes specific structures as dates, quantities and personal NP's

  [[PN[3]: Leonardo Da Vinci] <SUBJ [V:dibujar <**qtOBJ [ENTITY]** <DATEen 1492]]

  ([[PN: Leonardo Da Vinci] <SUBJ [V:draw <**qtOBJ [ENTITY]** <DATEin 1492]])

Currently, SpQA identifies six different semantic targets: PERSON, ENTITY, QUANT, TIME, PLACE, MANNER. To identify them, the parser uses the linguistic information encoded in the wh-words.

- PERSON: when the target is human.

  ¿Quién era el presidente de Francia durante las pruebas de armas nucleares en el Pacífico Sur?

  (Who was the president of France during the tests of nuclear weapons in the South Pacific?)

- ENTITY: when the target is no human.

  ¿Qué fue levantado el 13 de agosto de 1961?

  (What was built the 13th of August 1961?)

- QUANT: when the target is a quantity.

  ¿Cuántos goles se marcaron en total en el Mundial de Fútbol de 1982?

  (How many goals were scored in total in the World Cup of 1982?)

- TIME: when the target is related with time (a date, time, etc).

  ¿Cuándo se firmó el Tratado de Maastricht?

  (When was the Maastricht Treaty signed?)

- PLACE: when the target is a location.

  ¿Dónde se celebraron los JJ.OO. de 1992?

  (Where were celebrated the Olympic Games of 1992?)

- MANNER: when the target is a process, a description or an explanation.

  ¿Cómo actúa la hormona del crecimiento?

  (How does growth hormone work?)

When the question has a more complex interrogative constituent like

*¿***Cuántos kilos de anchoas** *capturó la flota del Cantábrico durante 1994?*

(How many kilos of anchovies did the fleet of the Cantabric fish in 1994?)

SpQA identifies the semantic target with the nucleus of the interrogative constituent:

[[[[N: flota] <PREPde [PN: Cantábrico]] <DET la] <SUBJ [V:capturar <**qtOBJquant [[N: kilos]** <**PREPde [N: anchoas]]** <DATEdurante 1994]]

([[[[N: fleet] <PREPof [PN: Cantabric]] <DET the] <SUBJ [V:fish <**qtOBJquant [[N: kilos]** <**PREPof [N: anchovies]]** <DATEin 1994]])

4 Question Parsing Evaluation

We are interested in a comparative evaluation of SpQA against other Spanish general parsers. However, building the methodology and the necessary data for parsing evaluation is a very complex and hard task (especially if the parsers have different frameworks, like in our case). For this reason, for our evaluation we have used the same data and evaluation methodology of Gayo (2011), applying them to SpQA and comparing our results with those of Gayo (2011) for DepPattern, Txala and Hispal.

In this section, we present first the three Spanish parsers used for the comparative evaluation of SpQA: Txala, Hispal and DepPattern. Then, we explain in detail the comparative evaluation method taken from Gayo (2011).

4.1 Spanish Parsers for the Comparative Evaluation

TXALA is the Spanish parser in the suite *Freeling*[4] (Padró et al., 2010). It can be downloaded for free (as a part of *Freeling*) and it is also available on-line. It offers dependency parsing with functional labeling.

HISPAL is the Spanish parser of the VISL[5] project. It is only available for use on-line, but it allows the uploading of files for analysis with a maximum of 2 Mb. It performs constituent parsing with functional labeling in the Constraint Grammar framework.

DEPPATTERN is the Spanish parser in the suite *DepPattern Toolkit*[6] (Gamallo and Sánchez, 2009). It can be downloaded for free and it is also available on-line. It offers dependency parsing with functional labeling.

[3] PN = Proper Noun

[4] http://nlp.lsi.upc.edu/freeling/

[5] http://beta.visl.sdu.dk/

[6] http://gramatica.usc.es/pln/tools/deppattern.html

4.2 Evaluation Methodology

For the comparative evaluation of SpQA, we have used the parser evaluation methodology presented in Gayo (2011). We applied the metrics of PARSEVAL scheme (Black et al., 1991) to measure two variables in question analysis: constituent recognition and constituent labeling. For each variable, we measure

- **Precision**: number of correct constituents (constituents in the gold standard) in parser output divided by number of constituents in the parser output.
- **Recall**: number of correct constituents (constituents in the gold standard) in parser output divided by the number of constituents in the gold standard.
- **F1 score**.

We applied these two variables to constituents in general (all the constituents in the sentence) and to the interrogative constituent in particular (for the importance of this element in QA systems). To make possible the comparison of SpQA with the results showed in Gayo (2011), we also used the same testing corpus of questions and the same gold standard.

4.2.1 The Testing Corpus

The corpus is made up of 100 questions extracted from monolingual Spanish sets of CLEF[7] 2004, 2006 and 2007. All the examples in the testing corpus are wh- questions. Questions were selected from CLEF sets according to their syntactic structure. The idea was to choose questions that presented a variety of syntactic structures, like different interrogative constituents, subordinated clauses, dates or named entities.

4.2.2 The Gold Standard

The gold standard is made up of the 100 questions of the testing corpus analyzed manually by one person. The analysis consists of the identification of the main syntactic structure (constituents in the sentence): verb and arguments/adjuncts, labeled with their syntactic function.

¿Qué robaba el oso Yogui?
What did Yogi Bear steal?
3 constituents:
Verb: *robaba* (did...steal)
Interrogative Direct Object: *Qué* (what)
Subject: *el oso Yogui* (Yogi Bear)

To minimize possible differences between parsers caused by their different frameworks, some linguistic decisions were taken in the annotation. These decisions tried to simplify as much as possible the syntactic analysis. For example, we only consider six syntactic labels: subject (S), direct object (O), indirect object (IO), predicative (PR), adjunct (CC; bounded or unbounded) and modifier (MOD); we analyze the verbal phrase always as one constituent (even if it was a complex unit: *ha sido premiado*, has been awarded); we do not compute as constituents functional clitics as *lo* (direct object clitic) or *se* (impersonal clitic); etc.

4.3 Parsers Output Analysis

For Txala, Hispal and DepPattern we use directly the data of Gayo (2011). For SpQA, we analyzed the testing corpus with the parser and we extracted:

- Number of constituents recognized: total number of constituents in the parser output.
- Identification of constituents: number of correct and incorrect constituents (compared with the gold standard) in the parser output.
- Labeling: number of correct and incorrect labeled constituents (compared with the gold standard) in the parser output.

5 Results

We show first the results concerning question constituents in general. Then, the results related to the interrogative constituents in particular (identification and labeling for both).

5.1 Question Constituents

We can see the results of general constituent recognition in Table 4.

	Hispal	**Txala**	**DepPatt.**	**SpQA**
precision	86.9	89.9	88.8	91.2
recall	88.7	93.3	83.6	93.6
F-score	87.8	91.6	86.1	92.4

Table 3: Constituent recognition.

The four parsers have good results: around or over 90. SpQA has the best results, although they are very close to those of Txala.

[7] http://www.clef-campaign.org/

In general constituent labeling, we have the next results:

	Hispal	Txala	DepPatt.	SpQA
precision	72.5	73.9	56.1	94.5
recall	64.3	69.0	46.9	88.5
F-score	68.2	71.3	51.1	91.4

Table 4: Constituent labeling.

Again SpQA has the best results. However, for this variable there is a clear distance between SpQA and the other three parsers. Whereas the accuracies of Txala, Hispal and DepPattern drop significantly in this task (comparing their results with Table 4), SpQA maintains its performance (only the recall is a bit lower).

So, as we can see, SpQA shows very close results in general constituent recognition and labeling.

5.2 Interrogative Constituent

Table 5 shows the results for interrogative constituent recognition:

	Hispal	Txala	DepPatt.	SpQA
precision	96.1	100.0	90.0	99.0
recall	98.0	100.0	90.0	99.0
F-score	97.0	100.0	90.0	99.0

Table 5: Interrogative constituent recognition.

Again, the four parsers have good results, all over 90. Txala has the best accuracy, followed very closely by SpQA.

The reason that SpQA does not achieve an accuracy of 100 is simple: the parser fails in the recognition of one of the sentences as a question, due to structural syntactic reasons (the question has a syntactic order that is not in the grammar). As a consequence, with the current architecture of the system, this causes it to fail in the recognition of the interrogative constituent. However, the important thing to note is that the problem is not in the recognition of the interrogative and it can be easily solved.

Concerning labeling, these are the results:

We can see again a substantial difference in parser accuracies between recognition and labeling. Hispal, Txala and DepPattern especially, have worse results again, whereas SpQA keeps its accuracy.

	Hispal	Txala	DepPatt.	SpQA
precision	52.0	62.0	25.0	94.9
recall	53.0	62.0	25.0	94.0
F-score	52.5	62.0	25.0	94.5

Table 6: Interrogative constituent labeling.

The accuracy in interrogative constituent labeling is even lower than in general constituent labeling (Table 5) for Hispal, Txala and DepPattern. From accuracies around 70, Hispal and Txala drop to numbers around 50 and 60, respectively; DepPattern falls from 51 to 25.

On the other hand, SpQA still maintains its performance, and, contrary to the other two parsers, it has even better results labeling the interrogative constituent (94%) than in general labeling (91%).

6 Conclusions and Future Work

Question processing is a crucial step in QA systems. In this processing, syntactic analysis of questions plays an important role.

For this task, we have presented SpQA, a parser focused on question analysis in Spanish. Currently, the system recognizes and labels all the constituents in the question. In addition, it identifies the syntactic and semantic target of the questions, as well as dates, proper nouns and quantities.

Compared to three freely available Spanish parsers, Hispal, Txala and DepPattern, SpQA shows the best results in four tasks: recognition and labeling of general constituents and recognition and labeling of the interrogative constituent. Besides this, whereas Hispal, Txala and DepPattern show a considerable difference between their accuracies in constituent recognition and labeling (general and for the interrogative constituent), SpQA keeps its accuracy, which is always over 90.

Future work concerns syntax and semantics aspects of SpQA. First, we have to make the grammar more complete to cover all possible syntactic structures of Spanish questions. Then, it will be necessary to concentrate on semantic aspects of questions, especially on the aspects related to target identification.

References

Jordi Atserias, Elisabet Comelles, and Aingeru Mayor. 2005. Txala un analizador libre de dependencias

para el castellano. *Procesamiento del Lenguaje Natural*, (35):455–456, September.

Eckhard Bick. 2006. A Constraint Grammar-based Parser for Spanish. In *Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology*.

Ezra Black, Steven P. Abney, D. Flickenger, Claudia Gdaniec, Ralph Grishman, P. Harrison, Donald Hindle, Robert Ingria, Frederick Jelinek, Judith L. Klavans, Mark Liberman, Mitchell P. Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzalkowski. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *North American Chapter of the Association for Computational Linguistics*.

Gracinda Carvalho, David Martins de Matos, and Vitor Rocio. 2010. Improving IdSay: A Characterization of Strengths and Weaknesses in Question Answering Systems for Portuguese. In Thiago Alexandre Salgueiro Pardo, António Branco, Aldebaro Klautau, Renata Vieira, and Vera Lúcia Strube de Lima, editors, *PROPOR*, volume 6001 of *Lecture Notes in Computer Science*, pages 1–10. Springer.

Jennifer Foster. 2010. "cba to check the spelling" Investigating Parser Performance on Discussion Forum Posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 381–384, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pablo Gamallo and Isaac González. 2009. Una gramática de dependencias basada en patrones de etiquetas. *Procesamiento del Lenguaje Natural*, (43):315–323.

Iria Gayo. 2011. Análisis de preguntas para Búsqueda de Respuestas: evaluación de tres parsers del español (Question Analysis for QA: Evaluation of three Spanish Parsers). In *Proceedings of SEPLN'11 (to appear)*.

Daniel Gildea. 2001. Corpus Variation and Parser Performance. In Lillian Lee and Donna Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202.

Ulf Hermjakob. 2001. Parsing and Question Classification for Question Answering. In *Proceedings of the workshop on Open-domain question answering - Volume 12*, ODQA '01, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cornelis H. A. Koster, Marc Seutter, and Olaf Seibert. 2007. Parsing the Medline Corpus. In *Proceedings RANLP 2007*, pages 325–329.

Marina Lloberes, Irene Castellón, and Lluís Padró. 2010. Spanish FreeLing Dependency Grammar. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

David Mcclosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and Self-training for Parser Adaptation. *ACL-COLING*.

D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu, and O. Bolohan. 2002. Lcc Tools for Question Answering. In Voorhees and Buckland, editors, *Proceedings of the 11th Text REtrieval Conference (TREC-2002)*, NIST, Gaithersburg.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five Years of Open-source Language Processing Tools. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for Accurate Deterministic Question Parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA, October. Association for Computational Linguistics.

José Luis Vicedo. 2004. La búsqueda de respuestas: Estado actual y perspectivas de futuro. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 8(22):37–56.