Assigning Wh-Questions to Verbal Arguments in a Corpus of Simplified Texts

Magali Sanches Duran[1], Marcelo Adriano Amâncio[1], Sandra Maria Aluísio[1]

[1] Center of Computational Linguistics (NILC) – Department of Computer Science
University of São Paulo – São Carlos-SP Brazil
magali.duran@uol.com.br, marcelousp@gmail.com, sandra@icmc.usp.br

Abstract. This paper reports a new annotation task that assigns wh-question labels to verbal arguments in a corpus of simplified texts in Portuguese. The aim is to provide a training corpus to a wider project called PorSimples, which has among its goals eliciting sense relations between verbs and their arguments through the exhibition of question words such as *who, what, which, when, where, why, how, how much, how many how long, how often* and *what for*. The annotation task involves recognizing segments that constitute answers to questions made to the verbs and deciding which question label is suitably answered by each segment. Results of such a semantic layer of annotation may be used, in addition, to identify adjunct semantic roles and multi-word expressions with specific adverbial syntactic roles.

Keywords. semantic annotation. wh-question labels. semantic role labeling.

1. Introduction

This annotation work has arisen from a wider project called PorSimples [1, 2, 3], which aims, among other goals, to elicit relations between verbs and other elements of a sentence through the exhibition of question words. Such initiative has a pedagogical purpose: to support users that can hardly interpret a text.

There are two sub-tasks involved in this annotation work: one of them is recognizing arguments boundaries and the other is assigning the wh-question answered by such arguments.

In this paper, the expression "verbal arguments" is employed to make reference to both: arguments predicted by verb senses and adjuncts that modify verb senses adding information about circumstances of time, locative, manner, purpose, cause and quantity.

We describe here the pilot test of the annotation task, which had the purpose of evaluating the reproducibility of the task and evidencing features required from an annotation tool for such task.

The corpus chosen for this work consists of simplified version of 154 texts extracted from newspapers [4]. These simplified texts were downloaded from Portal of Parallel Corpora of Simplified Corpus[1] .

There are two main reasons considered here to annotate a simplified corpus:

1. Simplified texts consist of active sentences, have no relative clauses, no appositions and have few coordinate and subordinate clauses; features which made them less exposed to automatic parsing errors. This is intended to ensure a better performance in the automatic steps of pre-annotation process as well as to provide a better input for the future steps of learning rules.

[1] http://caravelas.icmc.usp.br/portal/index.php

2. Simplification rules used to generate the texts of the corpus [5] did not produce changes relating to adjuncts, that is, they do not include losses of relevant material for the intended annotation.

This corpus has been previously annotated by the parser Palavras [6], but syntactic annotation has not been submitted to human correction.

Besides its first purpose of serving as a training corpus for the machine learning approach of the assignment of questions to verbal arguments, the resulting annotated corpus may be useful to 1) map semantic role labels, as there is correspondence between wh-questions and adjunct semantic roles like *time* (when), *place* (where), *quantity* (how much and how many), *manner* (how), *purpose* (what for) and *cause* (why); 2) provide data to improve parsers with fine-grained adverbial syntactic roles.

The annotated corpus resulting from this work will be made publicly available and shall benefit other applications related to automatic processing of Portuguese.

Section 2 discusses the problems related to assigning questions. Section 3 presents the approach used to assign wh-questions to verbal arguments using parser Palavras [6] to pre-annotate arguments boundaries and the annotation tool MMAX2 [7, 8]. Further, in Section 4, we report the evaluation of task reproducibility through Kappa Statistics [9].

2. Asking questions answered by verbal arguments

In our pilot study we worked with a list of 43 defined question labels, among which 26 relates to adjuncts and 17 relates to predictable arguments like “Quem?”, “O quê?”, “Qual?”(*Who, What, Which, Whose*) and combined forms with prepositions.

Hagège’s [10] proposal to identify time expressions inspired the inclusion of two questions in our previous list of questions: “Quanto tempo?” (=*How long?*) and “Com que frequência?” (=*How often?*) These two questions allow distinguishing temporal adjuncts that expresses frequency and duration aspects of time.

Question labels were organized into two levels: one for types and another for subtypes. Types relates to main questions (“o quê?”, “quem?”, “qual?”, “como?”, “onde?”, “para quê?”, “por quê?”, “quando?”and “quanto?”). Each type has its own subtypes. For the type “quando?”(=*When*), for example, which conveys time meaning, there are nine subtypes: “quando?” “de quando?” “desde quando?”, “até quando?”, “para quando?” “com que frequência?”, “quanto tempo?” “de quanto tempo?” “por quanto tempo?”.

Questions may be divided into three groups: 1) questions answered by subjects and direct objects (“quem?”, “o quê?”, “qual?”, “quais?”); 2) questions answered by indirect objects (“de quem?”, “para quem?”, “de quê?”, “com o quê?”, “sobre o quê” etc.) and 3) questions answered by adverbials (“onde?”, “quando?”, “quanto?”, “por quê?”, “como?”, “para quê?”).

Depending on the verb, there is ambiguity between the questions answered by the subject and by the direct object. In such case, question position is relevant. For example, “quem” before the verb will be related to the subject and “quem” after the verb will be related do the direct object: “João ama Maria” “Quem ama?” (João) “Ama quem?” (Maria). To face this problem, different labels had to be defined: “quem-direita” and “quem-esquerda”, the same for “o quê”, “qual” and “quais”.

Sometimes it is difficult to decide between “quem” and “o quê”, as for example in: “O carro atropelou o menino”. It seems to us that the better question is “Quem atropelou?”, in spite of “carro” being an inanimate noun. The annotation task will enable us to know which verbs require “quem” as subject or object, which verbs accepts both “quem” and “o quê” and, in this case, which features may be used to choose the suitable question.

Depending on the verb, there is also ambiguity between indirect objects and adverbials. In the example: “Ele pensa em silêncio”, the argument “em silêncio” is not an indirect object of the verb “pensar”, in spite of such verb allowing an indirect object introduced by the preposition “em” like in “Ele pensa em amizade”. To solve this problem, it is necessary to identify multiword expressions that convey adverbial sense, like “em silêncio” which is an adverbial expression of manner. The challenge is to decide whether the preposition belongs to the verb or to the adverbial. The annotated corpus will enable us to identify verbs and their possible complements introduced by prepositions, the question label distinguishing indirect objects from adverbials.

Another possible ambiguity exists between adverbials introduced by the same preposition. The preposition “em”, for example, may introduce a place: “Ele trabalha *em casa*” (“onde?”); a time: “Ele chega em uma semana” (“quando?”); a manner: “Ele que falar *em particular*” (“como?”); a cause “Ele não foi trabalhar *em função das enchentes*” (“por quê?”); a purpose: “Ele trabalha *em prol das crianças carentes*” (“para quê?”). Many of these ambiguities may be solved by identifying multiword expressions.

3. Our approach of annotation

The tool used for the annotation task in our pilot study was MMAX2[2] [7, 8], a free open source software. MMAX2 framework consists of two interfaces: one for configuration and one for annotation. Our intention was to realize, through the experience, which features should be requested from an annotation tool for our task.

MMAX2 allows combining different layers of annotation (multi-level annotation). It presupposes that every annotation consists of segments, called *markables*, that carry attributes and relations to each other. Therefore, MMAX2 tool was conceived to allow *markables* creation at different levels; each level of annotation is stand-off and data are stored in XML format.

The annotation process has been organized into three steps. In the first one, the main verbs or verbal phrases are identified. In the second step the boundaries of each argument are marked. Finally, the third step is to assign a question label to each argument.

To accomplish this process using MMAX2, we created four *markables*:

- *markable Tokens* – to annotate tokens;
- *markable Sentences*– to annotate sentences’ boundaries;
- *markable Verbs* – to annotate main verbs or verbal phrase of each sentence;
- *markable Arguments* – to annotate arguments’ boundaries.

All the four markables have been automatically annotated or automatically pre-annotated using information derived from a syntactic tree analysis provided by Parser Palavras [6].

[2] mmax2.sourceforge.net/

Using the parsed tree in TigerXML format, the tokenization was automatically performed by selecting the label “word” from *terminals* sub-tree, which represents each one of the sentence tokens. Sentence segmentation was achieved by simply selecting the label “s”.

By the other hand, segmentation of markables *Verbs* and *Arguments* were semi-automatic as they required human revision. *Verbs* segmentation considered any sub-tree, direct daughter of the principal node that was labeled by parser analysis as Verbal Phrase (VP). In coordinated sentences, there were two (or more) main verbs annotated, whereas subordinated sentences were annotated as one of the arguments of the main verb. *Arguments* segmentation considered all the other sub-trees, that is, direct daughters of the principal node that were not labeled as Verbal Phrase.

Examples of syntactic functions annotated as arguments include: subject, direct object and prepositional complements. Hand correction of arguments segmentation excluded connectives, discourse markers and other segments that do not answer wh-questions.

In the experiment for evaluating the reproducibility of the categorization of wh-questions (Section 4), verbs were highlighted by green color and parenthesis delimitation, whereas arguments were highlighted by blue color and square brackets, as shown in Figure 1.

[*Por causa da falta de segurança*], (**transferiram**) [*os cerca de 300 alunos*] [*para a escola Silveira de Souza*].

(Due to security risks, about 300 elevens have been moved to Silveira de Souza school.)

Figure 1: Example of a sentence after application of styles to the markables Verb and Arguments.

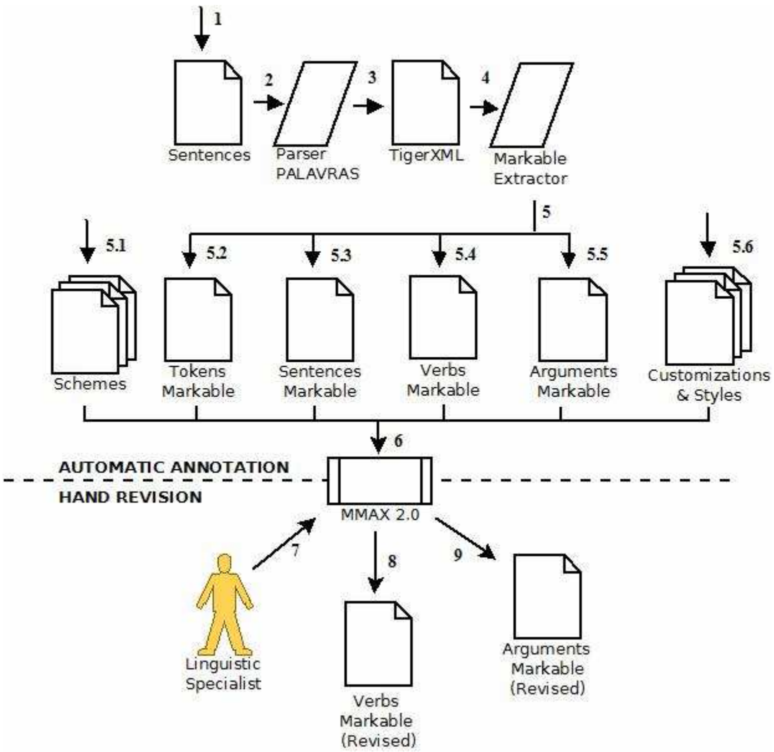

Figure 2 Annotation Process – Automatic and Semi-automatic Steps

Figure 2 shows the whole process of configuring MMAX2 (first and second steps of annotation process). A dotted line separates the figure into two areas: one related to automatic annotation and the other related to the human revision task.

The third step of the process was the only one fully hand annotated. Annotators were asked to choose a question label, from a pre-defined list of attributes, which was properly answered by the argument being annotated.

The advantage of organizing the process in steps is to divide the main task into simpler tasks, contributing to reduce probability of errors.

This experience with MMAX2 in the pilot study allowed us to observe desirable and undesirable features of an annotation tool for our task. To guide our choice, we elaborated the following list of requirements:

- labels and attributes edition during annotation task
- multi-level annotation;
- multi-level search engine;
- annotation on parse trees;
- comments edition during annotation;
- whole visualization of segments already labeled
- configuration of user's rights to read and edit labels
- sub-specification of labels
- graphical interface
- annotation of sentences that must be discarded or reexamined;
- easy label selection

4. Evaluating WH-Questions Annotation Task

A Question Annotation Manual[3] was drawn up to provide support to the annotation task. Each question label is illustrated by sentences that contain appropriate answers to it.

Aiming to test the reproducibility of the task, we gave a copy of this guide to seven annotators, all of them taking part of a post-graduation program on Computer Science. They had fifteen minutes to read the four pages of the guide and fifteen minutes to ask questions about it. After, they were shown an example of how to deal with the annotation tool, MMAX2, and then they started the annotation task that lasted not longer than an hour. This task took place in a laboratory and the annotators were required to assign questions to 75 arguments grouped into 25 sentences.

Kappa inter-annotator agreement shows a result of 0.78, indicating the task is reproducible. This result is very good considering that annotators were not linguistics specialists and training period was very short.

Inter-annotators disagreements provided new evidence on some possible ambiguities in question annotation task. These ambiguities were analyzed and some adjustments were made on the Manual.

[3] http://caravelas.icmc.usp.br/MakeExplicit

There was occurrence of arguments that allow two different question labels. In the sentence “Ele vai pedi-la em casamento na formatura”, the argument “na formatura” is an event, and an event takes place in a given local and time, so both questions *where* and *when* might be answered by it. Therefore, we took an arbitrary decision, defining at the Manual that the question *where* should prevail for labeling events.

Another case is “Os votos chegaram pelo telefone e pela internet”. In this case, the Manual established that the proper question to be assigned to semantic roles referring to means of communication and means of transport must be *Como?* So, “pelo telefone” (=by telephone) and “pela internet” (=by internet) should be annotated with “Como?” (=*How?*). Nevertheless, some annotators chose “Por onde?” (*by where?*). This is a point to be emphasized in annotators training.

5. Conclusion and Future Work

This work should be easier if we had a parser to provide a fine-grained labeling of adverbials and a corpus of Portuguese annotated with semantic roles. As it is not the case, the resulting annotated corpus with wh-question labels shall be useful, on its turn, as a starting point to semantic role labeling of Portuguese and as a data provider to improve parsers’ adverbial analyses.

Future work includes learning rules aiming at automatic question/role labeling. We expect the analysis of the fully annotated provides evidence on verbal features that influence question assignment, like type of transitivity, regency and voice.

Acknowledgments. Our thanks to FAPESP for supporting this work.

References

1. Aluísio, S. PorSimples: Simplification of Portuguese Texts for Digital Inclusion and Accessibility. 2st Joint MSR-FAPESP Workshop. Fapesp, 19/11/2008
2. Aluísio, S.; Specia, L.; Pardo, T; Maziero, E.; Caseli, H. M.; Fortes, R. (2008a) "A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems " In: Proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pp. 15-22.
3. Aluísio, S.; Specia, L.; Pardo, T; Maziero, E.; Fortes, R. (2008b) "Towards Brazilian Portuguese Automatic Text Simplification Systems. " In the Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), pp. 240-248.
4. Caseli, H.M.; Pereira, T.F., Specia, L.; Pardo, T.A.S.; Gasperin, C.; Aluísio, S.M.; (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In Alexander Gelbukh (ed), Advances in Computational Linguistics, Research in Computer Science, vol 41, pp. 59-70. 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), March 01–07, Mexico City.
5. Specia, L., Aluisio, S.M., Pardo, T.A.S.: Manual de Simplificação Sintática para o Português. Technical Report NILC-TR-08-06, São Carlos-SP. (2008)
6. Bick, E.: The Parsing System Palavras Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus, Denmark, Aarhus University Press. (2000)
7. Müller, C., Strube, M.: Multi-level Annotation of Linguistic Data with MMAX2. In: Braun, S., Kohn, K., Mukherjee J. (eds.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 197--214. Peter Lang: Frankfurt a.M., Germany (2006)
8. MMAX2: mmax2.sourceforge.net/
9. Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics, vol. 22, n. 2, pp. 249--254. (1996)
10. Hagège, C., Baptista, J., Mamede, N.: Proposta de Anotação e Normalização de Expressões Temporais da Categoria TEMPO para o HAREM II. http://www.linguateca.pt/aval_conjunta/HAREM/TEMPO_2008_02_18.pdf. (2007)