Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 43-50 recibido 11-01-11 revisado 13-02-11 aceptado 24-02-11

Automatic Question Categorization: a New Approach for Text Elaboration

Categorización automática de preguntas: un nuevo enfoque para elaboración de textos

Marcelo Adriano Amancio, Magali Sanches Duran, Sandra Maria Aluisio
Universidade de São Paulo ICMC-NILC São Carlos - SP, Brasil
marcelousp@gmail.com.br, magali.duran@uol.com.br, sandra@icmc.usp.br

Abstract. Text adaptation is a normal activity of teachers to facilitate reading comprehension of specific contents; the general approaches for it are Text Simplification and Text Elaboration (TE). TE aims at clarifying, explaining information and making connections explicit in texts. In this paper, we present a new approach for TE: an automatic question categorization system which assigns wh-question labels to verbal arguments in a sentence. For example, in "Mary danced yesterday." "Who?" is the label linking the verb "danced" to the argument "Mary" and "When?" links "danced" to the argument "yesterday". This annotation is similar to semantic role labeling, approached successfully via statistical language processing techniques. Specifically, we present experiments to build the system using a fine-grained question set in Portuguese language and address two key research questions: (1) Which machine-learning algorithm presents the best results? (2) Which problems this task presents and how to overcome them?
Keywords: Text Elaboration, Semantic Role Labeling, Wh-question labels.

Resumen: La adaptación de textos es una actividad normal de los profesores para facilitar la comprensión de la lectura de contenidos específicos. Los enfoques generales para esta actividad son la Simplificación de Textos y la Elaboración de Textos (ET). El objetivo de la elaboración de textos es esclarecer y explicar la información así como realizar las conexiones explícitas en éstos. En este artículo se presenta un nuevo enfoque para ET: un sistema automático de categorización de preguntas que asigna etiquetas de preguntas a los argumentos del verbo en la oración. Por ejemplo, en ¿María bailó ayer?, "Quién" es la etiqueta que enlaza el verbo "bailó" con el argumento "María" y "Cuándo" enlaza "bailó" con el argumento "ayer". Esta anotación es similar a la rotulación de roles semánticos, lo que constituye un enfoque que aplica con éxito técnicas de procesamiento estadístico del lenguaje. Específicamente se presentan experimentos para construir el sistema usando un conjunto amplio de preguntas en portugués y para responder las dos preguntas principales de esta investigación: (1) ¿Qué algoritmo de aprendizaje de máquina presenta mejores resultados? (2) ¿Qué problemas presenta esta tarea y cómo superarlos?
Palabras clave: Elaboración de texto, etiquetado de roles semánticos, rotulado de preguntas

1 Introduction

Text adaptation is a normal activity of teachers to facilitate reading comprehension of specific contents and also for language skills development (Burstein, 2009). It can benefit second-language learners and children learning to read texts of different genres. As well, text adaptation can benefit audiences with special needs, such as low-literacy readers, adults being alphabetized, people undertaking Distance Education (in which text understandability is of great importance), hearing-impaired people (who communicate to each other using sign languages and want to learn spoken languages, such as English or Portuguese), among others (Aluísio and Gasperin, 2010).

Studies in Text Adaptation try to answer two questions: What is modified? and How is it modified? With regard to the first question, researchers have investigated modifications at different linguistic levels: phonology, lexis, syntax, and discourse.

ISSN 1135-5948

As for the second question, there are two general approaches (or types) of text adaptation: Text Simplification (TS) and Text Elaboration (TE) (Young, 1999; Urano, 2000). The first can be defined as any task that reduces the complexity of a text (for example, lexical and syntactic complexity), while trying to preserve meaning and information (Siddarthan, 2003). As to TE, our focus in this work, it aims at clarifying and explaining information and making connections explicit in a text, for example, providing synonyms for words known to only a few speakers of a language or short definitions for complex concepts. TS and TE are strongly related; while TS enhances text readability, i.e., it makes the text easier to be read, TE is devoted to enhance text comprehensibility, i.e., it helps to increase easiness to understand concepts in a text. There are prominent studies on TE for the English language and a recent work for Portuguese; we relate them below. The Automated Text Adaptation Tool (Burstein, 2009; Burstein *et al.* 2007), for example, is a Natural Language Processing application for educational purposes, which is used by English language learners (ELLs) in content-area classrooms beyond elementary school. Since ELLs must learn the specialized, academic vocabulary which often includes low-frequency, more difficult words far beyond their English reading level, Text Adaptor includes an easier synonym adjacent to a difficult word and marginal notes (a kind of summary) translated into Spanish, besides other functionalities related to Text Simplification. Urano (2000) investigated the effects of lexical simplification and elaboration on sentence comprehension and incidental vocabulary acquisition by Japanese learners of English as a second language (L2). The modifications were carried out substituting unknown words (very low-frequency words) with high-frequency synonyms, and adding synonyms of the unknown words in apposition to them, respectively. The results of this study suggest that both lexical simplification and elaboration can improve learner comprehension at the sentence level. However, lexical elaboration resulted in incidental vocabulary acquisition, while simplification did not; and learners of higher proficiency benefited more from lexical elaboration in terms of the acquisition of word meanings. Instead of focusing on second language learners as the studies above, Watanabe *et al.* (2010) addressed low-literacy readers accessing Web pages and proposed a web content adaptation tool, named Educational Facilita. They used lexical elaboration (simple synonyms) and provided short definitions from Wikipedia to define named-entities (i.e., names of person, organization, location, among others) which appear in the text besides highlighting these entities. The set of named-entities used by the study was established by a taxonomy proposed in the evaluation contest of systems for recognizing named-entities in Portuguese (HAREM[1]). Moreover, they presented additional information about the highlighted named-entity, such as pictures for those entities of the person class.

In this paper, we present a new technique for TE intended to enable detailed reading of a text and accurate information extraction. Our ultimate goal is to build an automatic question categorization system which assigns wh-question labels to verbal arguments in a sentence. This initiative has a pedagogical purpose: to support users that can hardly comprehend a text, including children who are learning to read. For the best of our knowledge, this is a new task. Wh-question assignment task presented herein is a kind of semantic annotation which involves the subtasks of making delimitation of verbs and arguments, and linking verbs to their arguments through question labels.

Recent work in Natural Language Processing has shown the benefit of using statistical language processing techniques for the task of semantic role labeling (SRL), which is strongly related to our task. In this paper, we present several machine-learning experiments to build an automatic question categorization system which assigns wh-question labels to verbal arguments in a sentence. We use a fine-grained question set composed of 68 question labels in Portuguese language and address the following research questions: (1) Which machine-learning algorithm presents the best results?, (2) Which problems this task presents and how to overcome them?

In the remainder of this paper, we describe in detail the task of wh-question labeling assignment, giving emphasis to its relation with the task of SRL (Section 2), and then present the corpus, features, and question labels (Section 3), and the experiments performed to answer our research questions (Section 4). Section 5 summarizes our first contributions and indicates future work.

2 The Task of Wh-question Labeling Assignment

Wh-question assignment is a type of semantic annotation that links verbs to their arguments through wh-question labels such as who, what, which,

1 http://www.linguateca.pt/HAREM/

when, where, why, how, how much, how many how long, how often and what for. Figure 1 shows this annotation for the sentence "John went to Brazil last summer."

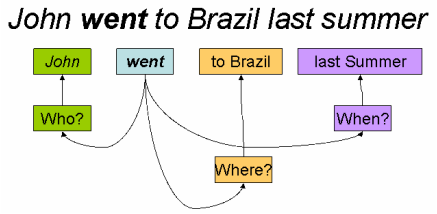

Figure.1: Example of question label assignment

In Figure 1, "Who?" is the question label that links the verb "went" to the argument "John". Similarly, "Where?" links the verb "went" to the argument "to Brazil" and "When?" links the same verb to the argument "last summer".

There is a commercial system that annotates actions and named-entities with wh-questions to support text mining[2]. Our task is different from the task performed by this system in the sense that we link verbs to all their arguments even if they are not named-entities. We use the term "argument" here in the same way it is used in the Propbank project (Palmer *et al*., 2005), i.e., on referring to both: arguments predicted by verb senses and adjuncts that modify verb senses adding information about circumstances of time (when), place (where), quantity (how much and how many), manner (how), purpose (what for), direction (in which direction) and cause (why).

Our task has two subtasks: 1) to set the boundaries of verbs and arguments and 2) to choose the question label that links properly the verb to each argument. In this paper we address the subtask 2. Some of the problems posed by this subtask are reported below in this section. Moreover, the use of a fine-grained set of questions has also some challenges such as those we present in Section 4.

In a pilot study we worked with a list of 43 defined question labels and conducted an experiment to determine the concordance of the question annotation task (Duran *et al.*, 2010). We created an annotation manual for this task and it was given to seven annotators to read it for about 30 minutes. After that, they took a time of about one hour to annotate 75 arguments occurring in 25 sentences. The resulting kappa was 0.78, indicating that the task is reproducible.

It is worth pointing out that in Portuguese, there are Wh-questions composed by prepositions and a question word, as for example, "De quem?" (*of who?), which explains the large number of question labels in our corpus (68). We have also created two labels "quem" (who): 1) "Quem?-DIR" related to Arg1 or Arg2 of Propbank role labels (syntactic role: direct object), and 2) Quem?-ESQ" related to Propbank's Arg0 or Arg1 (syntactic role: subject). The same was done for the labels "O quê?" (what), "Qual?" (which) and "Quais?" (which/plural), although in our corpus "Qual" and "Quais" only appeared at left. In the example (1) we show a sentence in active voice taken from our corpus, to illustrate the use of the labels "O que?-DIR" and "O que?-ESQ".

O Projeto_Rondon [o quê?-ESQ] é uma iniciativa do governo federal [o quê?-DIR] (The Rondon Project is an initiative of the Federal Government.) (1)

For the example (1), we have two questions: (i) What is an iniciative of the Federal Government? Answer: the Rondon project; (ii) The Rondon Project is what?* Answer: an iniciative of the Federal Government. Except for role labels associated to subject and direct object, question labels possess greater granularity than Propbank role labels. To illustrate this decision, we show below two examples, also taken from our corpus, of the set of eight question labels related to the semantic role of place: "onde?" (where?), "de onde?" (from where?), "aonde?" (to where?), "para onde?" (to where?), "por onde?" (by where?), "de onde?-filiação" (from where?-affiliation), "até onde?" (until where?), and "a partir de onde?" (from where?) (2-3).

A massa [o quê?-ESQ] que vem do polo Sul [de onde?] atinge os gaúchos [quem?-DIR] desde terça-feira [desde quando?]. (The mass that comes from the South Pole reaches the gaúchos since Tuesday.) (2)

EUA [quem?-ESQ] devem enviar mais 20 mil militares [o quê-DIR] ao Iraque [aonde?]. (U.S. should send more 20,000 soldiers to Iraq.) (3)

Besides the 16 labels presented above, there are 52 more[3], totalling 68 tags.

Depending on the verb, there is ambiguity between the questions answered by the subject and questions answered 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 to the direct object. To face this problem, different labels

2 http://www.cortex-intelligence.com/tech/

3 The complete tagset can be found in http://www.nilc.icmc.usp.br/porsimples/elatex

were defined: "Quem-direita" (Who-Right) and "Quem-esquerda" (Who-Left), the same for "O quê", "Qual" and "Quais".

The question answered by predicative is "Como?" (How), except for predicatives introduced by the verb "SER" (to be), which will be explained below. The question "Como?" is also assigned to adjuncts of manner. In order to allow future SRL, we created a question label "Como?-verbal" to distinguish predicative from adjuncts of manner that answer the question "Como?". Questions answered by indirect objects are: "De quem?" "Para quem?" "De quê?" "Com o quê?" "Sobre o quê?", etc. There is a lot of labels because in Portuguese the preposition that introduces indirect object is moved to the left of the wh-question. Questions answered by adverbials are: "Onde?" "Quando?" "Com que freqüência?" "Por quanto tempo?" "Quanto?" "Por quê?" "Como?" "Para quê?" "Em que direção?" and combinations of prepositions with the wh-questions "onde", "quando" and "quanto" (Por onde?, De onde?, De quando?, A quanto?, etc.).

Depending on the verb, there is ambiguity between indirect objects and adverbials. For example, in "Ele pensa em silêncio" (He thinks silently), "em silêncio" is not an indirect object of the verb "pensar", in spite of the fact that such verb admits an indirect object introduced by the preposition "em" like in "Ele pensa em amizade" (He thinks about friendship). 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. Another possible ambiguity exists between adverbials introduced by the same preposition. The preposition "em", for example, may introduce: (i) a place, "Ele trabalha em casa."/He works at home. ("Onde?"); (ii) a time, "Ele chega em uma semana."/He will arrive in one week. ("Quando?"); (iii) a manner, "Ele que falar em particular."/He wants to talk in private. ("Como?"); (iv) a cause "Ele não foi trabalhar em função das enchentes."/He did not go to work because the flooding. ("Por quê?"); a purpose: "Ele trabalha em prol das crianças carentes"/He works for the benefit of needy children. ("Para quê?"). Many of these ambiguities may be solved by identifying multiword expressions. In some sentences, there is information about somebody's institutional affiliation, introduced by verb "SER" (to be) like in (4). In these cases, we decided to assign a specific label "de onde?-filiação" (from where?-affiliation).

This tagset and annotated corpus were mapped to the tags of Propbank project (Palmer et al., 2005), i.e. numbered arguments Arg0, Arg1, Arg2, etc. and ArgMs (modifiers of the verb, such as manner (MNR), locative (LOC), temporal (TMP) and others), in order to, after manual revision, train classifiers for semantic role labeling of Portuguese sentences. Both corpora are available to download[4].

Cristiano_Zanuzo [quem?-ESQ] é de a corretora Renova [de onde?-filiação]. (Cristiano Zanuzo is the broker's Renova.) (4)

3 Corpus, Features and Question Labels

Our corpus is composed of 104 general news articles from Brazilian newspaper Zero Hora (ZH) which were manually simplified in the PorSimples project (Caseli *et al*. 2009). We have downloaded it from the Portal of Parallel Corpora of Simplified Corpus[5] and used a simplification version called "strong simplification".

The reasons for using a corpus of simplified texts were: (i) simplified texts consist of sentences in active voice, have no relative clauses, no appositions and have few coordinate and subordinate clauses, features which made them less exposed to automatic parsing errors, and (ii) the simplification rules used to generate the texts of the corpus did not produce changes related to adjuncts. This corpus was previously annotated by the parser Palavras (Bick, 2000), but the syntactic annotation was not revised. After the syntactic annotation, it was assigned 9820 question labels to their sentences, using the SALTO tool (*Burchardt et al*., 2006) and a tagging set with 68 different question labels. Table 1 shows a few statistics about the original and strongly simplified corpora.

Corpus	ZH original	**ZH strong**
Texts	104	**104**
Sentences	2184	**3329**
Words	46190	**43406**
Avg. words per text	444.1	**417,3**
Avg. words p. sentence	21.1	**13.0**

Table 1. Corpora statistics.

From a total of 3329 sentences annotated, 334 (9,1%) were flagged as "Wrong subcorpus" to be

[4] http://www.nilc.icmc.usp.br/porsimples/elatex

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

disregarded for the purpose of machine learning. The reasons for flagging a sentence as "Wrong subcorpus" are: parsing errors; errors of sentence splitting; titles of texts (not a sentence); and tokenization errors. Disregarding the "Wrong Subcorpus" flagged sentences, our corpus has 2295 sentences, 4771 verbs annotated (4151 simple verbs and 620 multiword verbs), and 9820 arguments annotated with question labels. In Figure 2 we observe that 3295 (33,55%) of annotated arguments are related to subject syntactic role ("What?-DIR" and "Who?-DIR") and 2966 (30,20%) are related to direct object syntactic role ("Who?-ESQ" and "What?-ESQ"). These were expected values, as subject and direct object are the more frequent verbal arguments. Relating to adjuncts, place, time and manner are the most frequent ones ("Where?", "When?", and "How?"). Indirect objects are well distributed, as the questions labels assigned to them are introduced by several different prepositions. They are included in "other labels", shown in Figure 2. It is worth mentioning that "Who" is the question more frequently answered by subjects (2120 "who" against 1175 "what") and "What?" is the question more frequently answered by direct objects (2753 "what"against 213 "who").

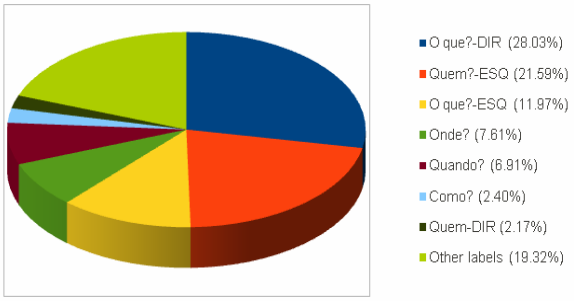

Figure 2:. The most frequent question labels assigned in our corpus.

As features, we are using mostly those proposed by Gildea & Jurafsky (2002), with some adaptation. Palmer et al. (2010) present some features introduced in recent SRL systems, besides the core features used by Gildea & Jurafsky (2002); we are also using some features from this work. Our feature set is composed by 23 features, presented below.

1) **Phrase type**: different question types tend to be realized by different syntactic categories. In general, Noun Phrases (NP) answer the questions "What?" and "Who?" while Prepositional Phrases (PP) answer questions with prepositions, such as "for what?", "of what?", "to where?", "in what?", "with whom?",. The parser Palavras, which annotated our corpus, has a large set of syntactic labels. For this feature we have used 12 higher level categories, such as adjectival phrases, adverbial phrases and clauses, besides NP and PP.
2) **Side** (or position): This feature indicates whether the constituint to be labeled occurs before (left) or after (right) the verb in focus. Therefore, there are two values: ESQ (left) and DIR (right) for this feature.
3) **Argument order**: This feature is an integer indicating the position of a constituent in the sequence of arguments for a given verb.
4) **Subcategorization of syntactic functions**: This feature refers to the set of a verb´s syntactic argument in the sentence. Since the parser Palavras has a large set of syntactic labels, this feature can have 26 values. as: direct object, indirect object, prepositional object, subject, predicator, utterance statement, subject complement, object complement, among others.
5) **Specific syntactic function**: This feature presents a subcategorization of the feature (4). For example, we have two types of direct object (DO), two types of indirect objects, two types of verbs (main verb and auxiliary verb),. This feature has 17 possible values.
6) **Question at the Left side?:** This boolean feature allows the identification of sentences without subject (a common phenomenon in Portuguese) or subjects at the right side of the verb.
7) **Number of arguments:** indicates the number of arguments of a sentence.
8) **Principal verb token**: an important lexical feature to determine the question type.
9) **First two Part Of Speech (POS) and Last POS of an argument**: These 3 features help to refine the type of NP involved, since the POS categories distinguish proper from common nouns and singular from plural nouns.
10) **First and Second tokens of an argument:** These features are used if the POS of the first and second tokens are from a closed class; for open class, they receive "—".
11) **Semantic values of the argument tokens:** For these features (in a total eight) it was used semantic categories (classes and subclasses) of the parser Palavras. Since the returned semantic classes and subclasses are lists, the first two elements were taken.

12) **Simple or Multiword verb:** The number of tokens of a Verb.
13) **Number of tokens of the argument:** This feature is an integer indicating the number of tokens of the argument.

4 Automatic Question Labeling for Portuguese: Experiments and Analysis

Section 4.1 shows our experiments with nominal classifiers, available in the Weka package (Witten and Frank, 2005): IBk, J48, JRip, SMO, and NaiveBayes. We also tried feature selection via InfoGainAttributeEval, available in the Weka package, in order to analyse which features are relevant; this is also described in Section 4.1. Section 4.2 tries to answer our second research question: (2) which problems this task presents and how to overcome them?

4.1 Machine-learning Methods and Feature Selection for Question Labeling

The Information Gain algorithm was chosen to rank the features because it is one of the most used methods. We started with 23 features and selected the 14 first ranked by the method. They are: (1) phrase type, (2) side, (5) Specific syntactic function, (4) Subcategorization of syntactic functions, (3) Argument order, (8) Principal verb token, (9) First POS of the argument, (9) Second POS of the argument, (9) Last POS of the argument, (10) First token of the argument, (10) Second token of the argument, (11) Specific Semantic value of the first token, (11) Generic Semantic value of the first token, (11) Specific Semantic value of the second token. The features eliminated have ranking values less than 0.34 whereas the first ranked has ranking value 1.39. We conducted this ranking step to reduce the data models size since we were not able to use Weka with all the features due to its memory limits.

Using the 14 best ranked features, we conducted our experiments using six machine-learning algorithms. SMO, SimpleLogistic (Maximum Entropy) and J48 had the best results of F-measure: 0.79, 0.78 and 0.74, respectively. They were followed by K-NN (k = 1[6]) with F=0.73. The worst results for F-measure were JRIP with F=0.72 and Naïve Bayes with F=0.71. For all algorithms, we used the 10-fold cross-validation procedure. All the methods performed better than a majority class (at LEFT and at RIGHT) baseline that is 41.84%. Considering F-Measure, we have found that the SMO is the better algorithm for our task. We performed the next three experiments using this algorithm. The results and discussion of these experiments are presented in the following sections.

Although we have used a simplifed sentences corpus what could improve the performance of our classifier, since that simplification reduces parsing errors (botteneck of Wh-question labeling) and sparcity of data, we had two challenges. The first one was a small corpus and the second a more detaild tagset than those used in semantic role labelling tasks. Even though having these shortcomings our results are similar than those of semantic role labeling taggers.

4.2 Problems of the Automatic Question Categorization task

We tested not distinguishing between "O quê" (What) and "Quem" (Who) as these question labels depend on the verb sense and on the animate/inanimate feature of argument nouns. For example, the verb "assassinar" (to murder) asks for a "Quem" (Who) question for both arguments placed at right and at left of the verb. The verb "influenciar" (to influence), on the other hand, admits both animate and inanimate subjects and objects. In this case, the decision between "Quem" (Who) and "O quê" (What) depends on semantic features of the argument nouns. As we have not annotation providing distinction between animate/inanimate nouns, there is no feature to support the learning of this. Without such distinction, the F-measure was of 0.84. The method performed better than a majority class baseline of 53.98%. In this experiment, we unified the labels "O quê?" and "Quem?", as well as all their respective prepositioned labels. For example, we mapped "O quê?" and "Quem?" to a label called "quê_quem". In the same way, we mapped "De quê?" and "De quem?" to a label called "De quê_quem?". The remaining label set after this unification was composed of 57 labels.

Another test we made was not distinguishing LEFT and RIGHT position of question labels related to the verb "SER" (to be). Our corpus of simplified texts has a great percentage of sentences with such verb linking two Noun Phrases (NP). This is a consequence of simplification process that gave origin to our corpus, since all the appositions were turned into single sentences using the verb "SER" (to be) like in: *"A dona de a casa é a vendedora Ruth_Miller_Loiola."*. (The housewife is the saleswoman Ruth Miller Loiola.). The NPs may

[6] In our work the best k was 1, since we had tags with very low frequency in our corpus.

change of place (at right and at left), without changing the sense of the sentence: "*A vendedora Ruth Miller Loiola é a dona de casa*.". (The saleswoman Ruth Miller Loiola is the housewife). The parsing identifies both NP and predicative. Besides that, "to be a housewife" is an attribute of the saleswoman Ruth Miller Loiola and not the opposite. Then, in spite of "a dona de casa" being at left, it comes at right of the verb in the question generated: "*Quem é a dona de casa?"* (Who is the housewife?). The other NP, "the saleswoman Ruth Miller Loiola", on its turn, comes at left of the verb in the question generated: "*The saleswoman Ruth Miller Loiola é o quê?"* (Ruth Miller Loiola is what?).

This test gave us a F-measure of 0.82, confirming our hypothesis that, when we have a NP at left and a NP at right it is difficult to decide which one predicates the other. It is important to note that here also the method performed better than a majority class baseline of 44.46%. Therefore, our challenge is to develop a feature that helps to recognize which is the entity being predicated and which is the attribute assigned to such entity. In this experiment, we unified the labels "Quem-ESQ" and "Quem-DIR", as well as "O quê-ESQ" and "O quê-DIR", remaining 66 labels.

We observed that 80.68% of the labels assigned concentrated on 10% of question labels (7 labels). After testing our hypotheses relating to the most frequent labels, we verified separately the precision of the 61 labels (90% of question labels) which correspond to 19.32% of the total labels assigned. Our aim was to find out whether the small number of occurrences affects machine learning or the features were enough strong to ensure a good performance. Our hypothesis in this fourth test was that accuracy would not be low, as the less frequent labels are almost always initiated by a preposition, and preposition is a good feature for our task. In this experiment, we removed the most frequent question labels: "O que-DIR", "Quem-ESQ", "O que-ESQ", "Onde", "Quando", "Como" e "Quem-DIR". After that, we have got 1897 instances from the original 9820 ones. The F-measure of 0.728 is a little low when compared to the F-measure of the complete label set but higher than its baseline that is 9.12%. Since that half of the 61 labels have at most 20 instances, enlarging our corpus can benefit the performance of them.

5 Conclusions and Future Work

In this paper, we have shown our initial exploring experiments towards creating an automatic question categorization system intended to enable detailed reading of a text and accurate information extraction. We have tested several machine-learning algorithms on this new task and also experimented with a feature selection algorithm, in order to select the most relevant features from a set of 23 features used for SRL, a related task.

Our experiments have shown that the SVM algorithm, with feature selection, achieved the best performance of F-measure on our task. Similarly to SRL, SVM and MaxEntropy are the best machine learning algorithms. We intend to continue implementing new features and using new machine-learning methods, such as reranking (Palmer et al., 2010), for finding the best overall labeling for all the arguments in the sentence.

We have found that the precision values for most frequent and less frequent question labels are very similar. This finding confirms our hypothesis that in spite of providing little instances to train our classifier, the less frequent question labels have better discriminative features than the most frequent labels. Therefore, we recommend keeping less frequent question labels since they allow more natural questions in Text Elaboration Systems.

To obtain a better performance, we intend to develop an external lexical resource that presents verbal restrictions: verbs that make and make not restrictions on the questions answered by subject (who or what) and verbs that make and make not restrictions on object (who or what). When the verb makes such restriction, the respective value will be provided (who or what). When the verb makes no restriction on the question label assigned to the subject and/or to the direct object, the challenge is to identify which nouns are animated (we do not have a lexical resource that provides such feature automatically).

Acknowledgments:

Our thanks to FAPESP for supporting this work.

References

Aluisio, S., Gasperin, C.: Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. In the Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, June, 2010, 46--53 (2010)

Bick, E.: The Parsing System Palavras Automatic Grammatical Analysis of Portuguese in a Con-

straint Grammar Framework. Aarhus, Denmark, Aarhus University Press (2000)

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S.: SALTO - A Versatile Multi-Level Annotation Tool. In: Proceedings of LREC-2006, Genoa, Italy (2006)

Burstein, J., Shore, J., Sabatini, J., Lee, Y.W., Ventura, M.: The automated text adaptation tool. In NAACL '07: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations on XX, 3--4 (2007)

Burstein, J.: Opportunities for Natural Language Processing Research in Education. In the *Proceedings of CICLing, 6--27* (2009)

Caseli, H.M., Pereira, T.F., Specia, L., Pardo, T.A.S., Gasperin, C., Aluísio, S.M.: 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, 10th Conference on Intelligent Text Processing and Computational Linguistics, 59--70 (2009)

Duran, M.S., Amancio, M.A., Aluísio, S.M.: Assigning Wh-Questions to Verbal Arguments: Annotation Tools Evaluation and Corpus. In: The Seventh Conference on International Language Resources and Evaluation (LREC), 2010, Valletta. CALZOLARI, N. et al. (eds) *Proceedings of the Seventh conference on International* Language Resources and Evaluation (LREC). Paris: ELRA, v. 1. p. 1445--1451 (2010)

Gildea. D., Jurafsky, D.: Automatic Labeling of Semantic Roles. *Computational Linguistics* Volume 28, Number 3, 1--45 (2002)

Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: A Corpus Annotated with Semantic Roles, *Computational Linguistics Journal*, 31:1 , 71--106 (2005)

Palmer, M., Gildea, D., Xue, N.: *Semantic Role Labeling*. Synthesis Lectures on Human Language Technology Series, ed. Graeme Hirst, Mogan & Claypoole (2010).

Siddharthan, A.: *Syntactic Simplification and Text Cohesion*. PhD thesis, University of Cambridge (2003)

Urano, K.: *Lexical simplification and elaboration: Sentence comprehension and incidental vocabulary acquisition*. Unpublished master's thesis, University of Hawai'i at Manoa, Honolulu (2000). Available at http://www.urano-ken.com/research/thesis.pdf

Watanabe, W.M., Candido Jr, A., Amancio, M.A., Oliveira, M., Pardo, T.A.S., Fortes, R.P.M., Aluísio, S.M.: Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. In the *Proceedings of the W4A-7th International Cross-Disciplinary Conference on Web Accessibility 2010*, (2010). Nova York: ACM Press, v. 1, 1--9 (2010)

Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*, 2nd Edition. Morgan Kaufmann, San Francisco (2005)

Young, D.N.: Linguistic simplification of SL reading material: Effective Instructional Practice? *The Modern Language Journal*, 83(3), 350--366 (1999)