Automatic Distractor Generation for Domain Specific Texts

Itziar Aldabe and Montse Maritxalar

IXA NLP Group,
University of the Basque Country, Spain
{itziar.aldabe,montse.maritxalar}@ehu.es
http://ixa.si.ehu.es/Ixa

Abstract. This paper presents a system which uses Natural Language Processing techniques to generate multiple-choice questions. The system implements different methods to find distractors semantically similar to the correct answer. For this task, a corpus-based approach is applied to measure similarities. The target language is Basque and the questions are used for learners' assessment in the science domain. In this article we present the results of an evaluation carried out with learners to measure the quality of the automatically generated distractors.

Keywords: NLP for educational purposes, semantic similarity, distractor.

1 Introduction

The generation of Multiple-Choice Questions (MCQ), one of the measures used for formative assessment, is difficult and time consuming. The implementation of a system capable of generating MCQs automatically would reduce time and effort and would offer the possibility of generating a great amount of questions easily. In our proposal, we use Natural Language Processing (NLP) techniques to construct MCQs integrated in didactic resources.

There are different NLP-based approaches which have proved that the automatic generation of multiple-choice questions is viable. Some of them focus on testing grammar knowledge for different languages, such as English [1] or Basque [2]. Others apply semantic features in order to test general English knowledge [3], [4] or knowledge of specific domains [5]. Our work is focused on the automatic generation of MCQ in a specific domain, i.e. science domain. The target language is Basque.

The objective is to offer experts a helping tool to create didactic resources. Human experts identified the meaningful terms (i.e. words) of a text which were to be the blanks of the MCQs. Then, the system applied semantic similarity measures and used different resources such as corpora and ontologies in the process of generating distractors[1]. The aim of this work is to study different

[1] The incorrect options of the MCQs.

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 27–38, 2010.

methods to automatically generate distractors of high quality. That is to say, distractors that correspond to the vocabulary studied by learners as part of the curricula.

As there must be only one possible answer among the options of each MCQ, experts had to discard those distractors that could form a correct answer. Our purpose was to evaluate the system itself by means of an evaluation in a real situation with learners. The results of a test exercise was used to measure the quality of the automatically generated distractors. The evidence provided by the results will be used to improve the methods we propose.

The paper is organised as follows: section 2 explains the scenario to generate and analyse the questions. The methods we have used to generate distractors are explained in section 3. Section 4 presents the experimental settings and section 5 shows the results obtained when evaluating the questions with learners. Finally, section 6 outlines some conclusions and future work.

2 Design of the Scenario

We designed an experiment in which most of the external factors which could have an influence on the evaluation process were controlled.

The multiple-choice questions were presented to learners together with the whole text. Each MCQ is a *stem* and a set of options. The stem is a sentence with a blank. Each blank presents different options, being the correct answer the *key* and the incorrect answers the *distractors*. Example 1 shows an example of MCQs in the context of use.

Example 1. Espazioan itzalkin erraldoi bat ezartzeak, bestalde, Lurrari ...6... egingo lioke, poluitu gabe. Siliziozko milioika disko ...7... bidaltzea da ikertzaileen ideia. Paketetan jaurtiko lirateke, eta, behin diskoak zabalduta, itzalkin-itxurako egitura handi bat osatuko lukete. Hori bai, ...8... handiegiak izango lituzke.[2]

6 a. babes	b. aterki	c. defentsa	d. itzala
7 a. unibertsora	b. izarrera	c. galaxiara	d. espaziora
8 a. kostu	b. prezio	c. eragozpen	d. zailtasun

The process of generating and analysing the questions consists of the following steps:

- Selection of the texts: experts on the generation of didactic resources selected the texts on an specific domain, taking into account the level of the learners and the length of the texts.
- Marking the blanks: the terms to be considered as keys had to be relevant within the text. The marking was carried out manually.
- Generation of distractors: for each stem and key selected in the previous step, distractors were generated.
- Choosing the distractors: experts had to verify that the automatically generated distractors could not fit the blank.

[2] 6 a. protection b. umbrella c. defense d. shadow.

- Evaluation with learners: each learner read the MCQs embedded in a text and chose the correct answer among 4 options.
- Item Analysis: based on learners' responses, an item analysis process was carried out to measure the quality of the distractors.

3 Distractor Generation

When generating distractors, the purpose is to find words which are similar enough to the key but which are incorrect in the context (to avoid the generation of more than one correct answer).

We wanted to generate questions to test the knowledge on a specific domain, i.e. the science domain. The implemented methods are based on *similarity measures*. For that, the system employs the *context* in which the key appears to obtain distractors which are related to the it.

3.1 Word Space: Latent Semantic Analysis

Similarity measures are usual in different NLP applications such us in generating distractors. Two main approaches are used: knowledge-based methods and corpus-based methods. In fact, some researches employ WordNet to measure semantic similarity [4], others use distributional information from the corpus [6] and finally, there are some studies which exploit both approaches [5].

Measuring similarity for minority languages has some limitations. The main difficulty when working with such languages is the lack of resources. In our case, the main knowledge-based resource for Basque [7] is not finished yet: the Basque WordNet[3] is not useful in terms of word coverage, as it has 16,000 less synsets for nouns than WordNet 3.0. As a consequence, we decided to choose as the starting point a corpus-based method to carry out the experiments. Nonetheless, we also used a knowledge-based approach to refine the distractor selection task (cf. Section 3.2).

The system uses context-words to compute the similarity deploying Latent Semantic Analysis (LSA). LSA is a theory and method for extracting and representing the meaning of words [8]. It has shown encouraging results in a number of NLP tasks such as Information Retrieval [9,10] and Word Sense Disambiguation [11]. It has also been applied in educational applications [8] and in the evaluation of synonym test questions [12].

Our system makes use of Infomap software [13]. This software uses a variant of LSA to learn vectors representing the meanings of words in a vector-space known as WordSpace. In our case, it indexes the documents in the corpora it processes and performs word to word semantic similarity computations based on the resulting model. As a result, the system extracts the words that best match a query according to the model.

Build Word Space and Search: As the MCQs we work with are focused on the science domain, the collected corpus consists of a collection of texts related

[3] Nouns are the most developed ones.

to science and technology [14]. The corpus is composed of two parts. For this work, we used the balanced part (3 million words) of the specialised corpus.

In the process of building the model, the matrix was created from the lemmatized corpus. To distinguish between the different categories of the lemmas, the matrix not only took into account the lemma but also its category. The matrix contains nouns, verbs, adjectives and adverbs.

Once we obtained the model based on the specialised corpus, we had to set the context to be searched. After testing different windows, we set the sentence as the context.

3.2 Methods for Distractor Generation

The words found in the model were the starting point to generate the distractors for which different methods can be applied. The baseline method (LSA method) is only based on the output of LSA. The rest of the methods combine the output of the model with additional information to improve the quality of the distractors.

LSA Method: The baseline system provides InfoMap with the whole sentence where the key appears. As candidate distractors the system offers the first words of the output which are not part of the sentence and match the same PoS. In addition, a generation process is applied to supply the distractors with the same inflection form as the key.

LSA & Semantics & Morphology: One of the constraints here is to avoid the possibility of learners' guessing the correct choice by means of semantic and morphology information.

Let us see as an example a question whose stem is *"Istripua izan ondoren, ... sendatu ninduen"* (After the accident, ... cured me) the key is *medikuak* (the doctor) and a candidate distractor is *ospitalak* (the hospital). Both words are related and belong to the same specific domain. Learners could discard *ospitalak* as the answer to the question because they know that the correct option has to be a person in the given sentence. The system tries to avoid this kind of guessing by means of semantic information. Therefore, applying this method, the system does not offer *ospitalak* as a candidate distractor.

The system uses two semantic resources:

a) Semantic features of common nouns obtained with a semiautomatic method [15]. The method uses semantic relationships between words, and it is based on the information extracted from an electronic monolingual dictionary. The extracted semantic features are animate, human, concrete etc. and are linked to the entries of the monolingual dictionary.
b) The Multilingual Central Repository (MCR) which integrates different local WordNets together with different ontologies [16]. Thanks to this integration, the Basque words acquire more semantic information to work with. In this approach, the system takes into account the properties of the Top Concept Ontology, the WordNet Domains and the Suggested Upper Merged Ontology (SUMO).

In a first step, this method obtains the same candidate distractors as the LSA method and then it proposes only those which share at least one semantic characteristic with the key. To do so, the system always tries to find firstly the entries in the monolingual dictionary. If they share any semantic feature, the candidate distractor is proposed; if not, the system searches the characteristics in MCR, which works with synsets. By contrast, the output of Infomap are words. In this approach, we have taken into account all the synsets of the words and checked if they share any characteristic. That is, if a candidate distractor and the key share any characteristic specified by the Top Concept Ontology, the WordNet Domains or SUMO, the candidate distractor is suggested.

One might think that after obtaining distractors which share at least one semantic characteristic with the key, the system does not need any extra information to ensure that they are valid distractors. However, working with all the senses of the words may yield not valid distractors in terms of semantics. Moreover, there are some cases in which two words share a semantic characteristic induced from MCR but which would not be suitable distractors because of their morphosyntax.

In the last step, the method takes the candidate distractors which share at least one semantic characteristic with the key and it takes into account morphosyntax.

Basque is an agglutinative language in which suffixes are added to the end of the words. Moreover, the combination of some morphemes and words is not possible. For instance, while the lemma "ospital" (hospital) and the morpheme "-ko" form the word "ospitaleko" (of the hospital), it is not possible to combine the lemma "mediku" (doctor) with the suffix "-ko", since "-ko" is only used to express the locative genitive case with inanimate words.

As the input text is previously analysed by a morphosyntactic analyser, the system distinguishes the lemma and the morphemes of the key. It identifies the case marker of the key and it generates the corresponding inflected word of each candidate distractor using the lemma of the distractor and the suffix of the key as basis.

Once distractors are generated, the system searches for any occurrence of the new inflected word in a corpus. If there is any occurrence, the generated word becomes a candidate distractor. The searching is carried out in a Basque newspaper corpus which is previously indexed using swish-e[4] to ensure a fast search.

That certain words do not appear in the corpus does not mean that they are incorrect. Those distractors that do appear in the corpus will be given preference over distractors of common usage.

In this step, the system tries to avoid candidate distractors which the learners would reject based on their incorrect morphology.

LSA & Specialised Dictionary: The third method combines the information offered by the model and the entries of an encyclopaedic dictionary of Science and Technology for Basque [17]. The dictionary comprises 23,000 basic concepts related to Science and Technology divided into 50 different topics.

[4] http://swish-e.org/

Based on the candidate distractors generated by the LSA method, the system searches in the dictionary the lemmas of the key and the distractors. If there is an appropriate entry for all of them, the candidate distractors which share the topic with the key in the encyclopaedic dictionary are given preference. Otherwise, the candidate distractors with an entry in the dictionary take preference in the selection process. In addition, those candidates which share any semantic characteristic (cf. 3.2) with the key have preference to be suitable distractors.

LSA & Knowledge-based Method: This method is a combination of corpus-based and knowledge-based approaches to measure the similarities. Similarity is computed in two rounds. First the system selects the candidate distractors based on LSA and then, a knowledge-base structure is used to refine the selection.

The knowledge-based approach [18] uses a graph-based method based on WordNet, where the concepts in the Lexical Knowledge Base (LKB) represent the node in the graph, and each relation between concepts is represented by an undirected edge[5]. Given an input piece of text, this approach ranks the concepts of the LKB according to the relationships among all content words. To do so, Personalized PageRank can be used over the whole LKB graph: given an input text, e.g. a sentence, the method extracts the list of the content nouns which have an entry in the dictionary and relates them to LKB concepts. As a result of the PageRank process every LKB concept receives a score. Therefore, the resulting Personalized PageRank vector can be seen as a measure of the structural relevance of LKB concepts in the presence of the input context. In our case, we use MCR 1.6 as the LKB and Basque WordNet as the dictionary.

The method is defined as follows: Firstly, the system obtains a ranked list of candidate distractors based on the LSA model. Secondly, the Personalized PageRank vector is obtained for the stem and the key. Thirdly, the system applies the graph-based method for 20 candidate distractors in the given stem. Finally, the similarities among vectors computed by the dot product are measured and a reordering of the candidate distractors is obtained.

4 Experimental Settings

A group of experts chose five articles from a web site[6] that provides current and updated information on Science and Technology in Basque. As selection criteria, they focused on the length of the texts as well as the appropriateness to the learners' level. First of all, the experts marked the blanks of the questions and then the distractors were automatically generated. To identify the best method to generate distractors, we designed different experiments where the system applied all the explained methods for each blank and text.

Blanks: Experts who work on the generation of learning material were asked to mark between 15 and 20 suitable terms in five texts to create multiple-choice

[5] The algorithm and needed resources are publicly available at http://ixa2.si.ehu.es/ukb

[6] www.zientzia.net

questions. The blanks were manually marked because the aim of the experiment was to evaluate the quality of the distractors in a real situation. When proceeding, the experts did not follow any particular guidelines but followed the usual procedure[7]. The obtained blanks were suitable in terms of the appropriateness of the science domain and the stems.

In all, 94 blanks were obtained. As we did not give them any extra information for the marking process, experts marked as keys nouns, verbs, adjectives and adverbs. However, our study from a computational point of view aimed at generating nouns and verbs. 69.14% of the obtained blanks were nouns and 15.95% verbs. This shows that the idea of working with nouns and verbs makes sense in a real situation.

Distractors: The distractors were automatically generated for each blank and method. In the case of the nouns, the four mentioned methods were applied and in the case of the verbs, two methods were applied: the LSA method and the LSA & specialised dictionary method[8].

As the distractor generation task is completely automatic, the possibility of generating distractors that are correct in the given context had to be considered. That is why before testing them with learners the distractors were manually checked.

For each method, we provided experts with the first four candidate distractors. We had foreseen to reject the questions which had less than three appropriate distractors. However, in all the cases three valid distractors were obtained. Just 0.95% of the distractors could be as suitable as the key and 3.25% were rejected as dubious.

For each selected text, we obtained four tests (corresponding to the four methods). Moreover, a fifth test was manually made by experts, who created three different distractors for each blank semantically close to the keys. It is important to point out that the experts did not have any information about the distractors obtained from the automatic process. Finally, the manually built tests were compared to the automatically generated ones.

Schools and Learners: Six different schools took part in the experiment. The exercise was presented to the learners as a testing task and the teachers were not familiar with the texts until they handed out the test to their students.

In all, 266 learners of Obligatory Secondary Education (second grade) participated in the evaluation. They had a maximum of 30 minutes to read and complete the test. The test was carried out in paper in order to avoid all noise[9]. 249 of the learners completed the test and their results were used to analyse the items (questions) (see section 5).

After finishing the testing, an external supervisor collected the results of the exercise in situ.

[7] In this step, the evaluation was blind.

[8] We did not apply the remaining methods because the verbs in the Basque WordNet need of manual revision.

[9] We did not want to evaluate the appropriateness of any computer assisted assessment.

5 Results

By means of this evaluation we intended to improve the automatic methods explained in section 3. The item analysis method was the basis of our evaluation.

The item analysis method reviews items qualitatively and statistically to identify problematic items. The difference between both reviews is that the qualitative method is based on experts knowledge and that the statistical analysis is conducted after the items have been given to students. This paper is focused on the statistical analysis. We have used R free software environment[10] for statistical computing and graphics of the learners' results.

5.1 Item Analysis and Distractor Evaluation

The analysis of item responses in a quantitative way provides descriptions of item characteristics and test score properties among others. There are two main theories to address the problem: Classical Test Theory (CTT) and Item Response Theory (IRT). Both statistical theories have been already used in the evaluation of the automatic generation of distractors [3], [5].

In this experiment, we explored item difficulty, item discrimination and distractors' evaluation based on CTT as [5] did. However, the results obtained by them and our results are not comparable since they test the MCQs separately and we test them within a text.

Item difficulty: The difficulty of an item can be described statistically as the proportion of students who can answer the item correctly. The higher the value of difficulty, the easier the item.

Item discrimination: a good item should be able to discriminate students with high scores from those with low scores. That is, an item is effective if those with high scores tend to answer it correctly and those with low scores tend to answer it incorrectly.

The point-biserial correlation is the correlation between the right/wrong scores that students receive on a given item and the total scores that the students receive when summing up their scores across the remaining items. A large point-biserial value indicates that students with high scores on the overall test are also answering the item correctly and that students with low scores on the overall test are answering the item incorrectly. The point-biserial correlation is a computationally simplified Pearson's r between the dichotomously scored item and the total score. In this approach, we use the corrected point-biserial correlation. That is, the item score is excluded from the total score before computing the correlation. This is important because the inclusion of the item score in the total score can artificially inflate the point-biserial value (due to correlation of the item score with itself).

There is an interaction between item discrimination and item difficulty. It is necessary to be aware of two principles: very easy or very difficult test items have little discrimination and items of moderate difficulty (60% to 80% answering

[10] http://www.r-project.org/

correctly) generally are more discriminating. Item difficulty and item discrimination measures are useful only to help to identify problematic items. Poor item statistics of the results should be put down to ineffective distractors.

Distractor evaluation: to detect poor distractors, the option-by-option responses of high-scored and low-scored learners groups were examined. To this purpose, two kind of results will be presented in the next section: the number of distractors never chosen by the learners and a graphical explanation of the used distractors.

5.2 Interpreting the Results of the Tests

Table 1 shows the average of item difficulty and item discrimination results obtained for all the items in a text. The table shows the results for the manually and automatically generated tests.

In the case of item difficulty, each row presents the item difficulty index together with the standard deviation, as well as the percentage of easy and difficult items. In this work, we have marked an item to be easy if more than 90% of the students answer it correctly. On the other hand, an item is defined as difficult when less than 30% of the students choose the correct answer. The results shown for the manually generated test are promising (near 0.5), and there is not significant differences among the automatic methods. All of them tend to obtain better results with the second text and tend to create easy items.

The results of item discrimination take into account the responses of the high-scoring and low-scoring students. The high-scoring group is the top 1/3 of the class, and the low-scoring group comprises students with test scores in the bottom 1/3 of the class. Regarding item discrimination, the corrected point-biserial index with its standard deviation as well as the percentage of items with negative values are shown in the table.

Even though all the results obtain a positive mean (a value of at least 0.2 is desirable), in 8 out of the 10 cases negative point-biserial indexes are obtained. These negative values represent the percentage of items correctly responded by a higher number of low-scored students than high-scored ones. To identify the reasons underlying these results we study the option-by-option responses of high-scored and low-scored groups. Such study led us to evaluate the distractors themselves.

Figure 1 shows in a graphic way the distribution of the distractors among the low-scored and high-scored groups. The x axe represents the number of low-scored students selecting a distractor and the y axe represents the number of high-scored ones. In this experiment we have studied the results related to 108 distractors, limiting the number of students per test to 20.

Regarding the number of different distractors, in the case of the manually generated distractors, 83 (76.85%) out of the 108 distractors were chosen. In the cases of the automatically generated distractors the results were 64 (59.26%) for the LSA method, 54 (50.00%) for the LSA & semantics & morphology method, 67 (62.04%) for the LSA and & encyclopaedic dictionary method and 60 (55.56%) for the LSA & knowledge-based method.

Table 1. Item Analysis

		Difficulty			Item Discrimination	
		Item Difficulty	Easy	Difficult	Corrected Point-biserial	Neg.
LSA	Text1	0.79 (±0.18)	29.41%	0.00%	0.22 (±0.25)	17.65%
	Text2	0.67 (±0.20)	5.26%	5.26%	0.10 (±0.21)	31.58%
LSA & sem. & morph.	Text1	0.83 (±0.15)	35.29%	0.00%	0.26 (±0.16)	0.00%
	Text2	0.71 (±0.21)	21.05%	10.53%	0.30 (±0.15)	5.26%
LSA & spec. dictionary	Text1	0.70 (±0.23)	17.65%	11.76%	0.22 (±0.14)	5.88%
	Text2	0.66 (±0.22)	5.26%	5.26%	0.22 (±0.35)	26.32%
LSA & Knowledge-based	Text1	0.76 (±0.22)	23.53%	11.76%	0.33 (±0.30)	11.76%
	Text2	0.68 (±0.19)	26.32%	15.79%	0.41 (±0.19)	0.00%
Manually generated	Text1	0.66 (±0.23)	0.00%	5.88%	0.13 (±0.21)	23.53%
	Text2	0.46 (±0.26)	0.00%	36.84%	0.14 (±0.21)	21.05%

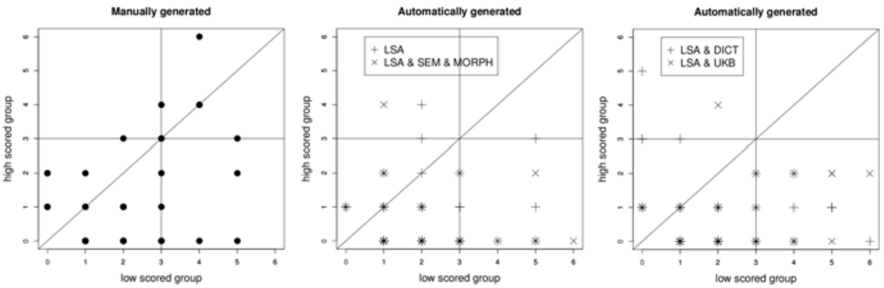

Fig. 1. Distractors Evaluation. * is used when both methods share the point.

Based only on the selected distractors, this last method gives the best results in relation to the percentage of distractors that discriminates positively among the low and high-scored groups: 90.00% (54 out of 60). The distractors obtained by the LSA & semantics & morphology method discriminated positively in 87.04% of the cases, the LSA & dictionary method in 79.10% of the cases, the LSA method in 76.56% of the cases and the manual method in 75.90% of the cases. In a graphic way, the distribution of the low-right side of the graphics can be interpreted as the set of good distractors.

The distribution of the high-left side of the graphics represents distractors that have to be revised because they confuse high-scored students and do not confuse low-scored learners. The reason could be that low-scored learners have not enough knowledge to be confused. Looking at the results of the methods, the LSA method tend to confuse more than the other methods (14.06%), followed by the manual method (12.05%), the LSA & dictionary method (11.94%), the LSA & semantics & morphology method (7.41%) and the LSA & knowledge-based method (6.67%).

It seems there is a relation between the number of the selected distractors and the percentage of discrimination: the lower the number of distractors, the higher the positive discrimination. However, the LSA method does not follow this assumption.

In order to improve the methods, we are planning to study in more depth the distractors that were never chosen. Moreover, it is also necessary to analyse on two other aspects: the domain nature and the part-of-speech of the keys. We must not forget that experts marked the blanks without being instructed. Therefore the blanks did not have to correspond with words related to the specific domain.

6 Conclusions and Future Work

The article presents a study about automatic distractor generation for domain specific texts. The system implements different methods to find distractors semantically similar to the key. It uses context-words to compute the similarity deploying LSA. We have used a balanced part of a specialised corpus to build the model. In the near future we will make use of the whole specialised corpus to model it.

In this approach, we have explored item difficulty, item discrimination and distractors' evaluation based on Classical Test Theory. The results shown for the manually generated test were promising, and there were not significant differences among the methods. The item discrimination measure led us to study the option-by-option responses of high-scored and low-scored groups and we finished the study with the evaluation of the distractors. Such evaluation gave us evidence to improve the methods regarding the domain nature and part-of-speech of the keys, and the need to enlarge the context when applying LSA. In addition, we are planning to test the distractors with more learners. Finally, the fact that the distractors tend to confuse high-scored learners, but not low-scored learners needs of deeper analysis.

In our opinion, working in a specific domain may improve the quality of the distractors so in the near future we will design new experiments with test exercises independent from the domain to compare the results with the ones obtained in the current study.

For future work we are also planning to use data mining techniques to identify the blanks of the text. Finally, reliability measures should also be considered in future research. Reliability tells us whether a test is likely to yield the same results if administered to the same group of test-takers multiple times. Another indication of reliability is that the test items should behave the same way with different populations of test-takers.

Acknowledgments. We would like to thank to the institution named "Ikastolen Elkartea" who has assigned good experts in order to work on the tasks of this work. This research is being supported by the Basque Government (SAIOTEK project, S-PE09UN36 and ETORTEK project, IE09-262).

References

1. Hoshino, A., Nakagawa, H.: Assisting cloze test making with a web application. In: Proceedings of SITE (Society for Information Technology and Teacher Eduation), San Antonio, U.S., pp. 2807–2814 (2007)

2. Aldabe, I., Lopez de Lacalle, M., Maritxalar, M., Martinez, E., Uria, L.: ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 584–594. Springer, Heidelberg (2006)
3. Sumita, E., Sugaya, F., Yamamota, S.: Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In: 2nd Workshop on Building Educational Applications Using NLP (2005)
4. Pino, J., Heilman, M., Eskenazi, M.: A Selection Strategy to Improve Cloze Question Quality. In: Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains (2008)
5. Mitkov, R., Ha, L.A., Varga, A., Rello, L.: Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. In: Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, pp. 49–56 (2009)
6. Smith, S., Kilgarriff, A., Sommers, S., Wen-liang, G., Guang-zhong, W.: Automatic Cloze Generation for English Proficiency Testing. In: Proceeding of LTTC conference, Taipei (2009)
7. Agirre, E., Ansa, O., Arregi, X., Arriola, J.M., Diaz de Ilarraza, A., Pociello, E., Uria, L.: Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis. In: Proceedings of the first International WordNet Conference, Mysore, India (2002)
8. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
9. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
10. Gliozzo, A.M., Giuliano, C., Strapparava, C.: Domain Kernels for Word Sense Disambiguation. In: 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005). University of Michigan, Ann Arbor (2005)
11. Schütze, H.: Automatic word sense discrimination. In: Computational Linguistics, vol. 24(1), pp. 97–124 (1998)
12. Turney, P.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
13. Dorow, B., Widdows, D.: Discovering corpus-specific word senses. In: Proceeding of EACL, Budapest (2003)
14. Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A.: ZT Corpus: Annotation and tools for Basque corpora. In: Copus Linguistics, Birmingham, UK (2007)
15. Diaz de Ilarraza, A., Mayor, A., Sarasola, K.: Semiautomatic labelling of semantic features. In: 19th International Conference on Computational Linguistics (2002)
16. Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P.: The MEANING Multilingual Central Repository. In: Proceedings of the Second International WordNet Conference-GWC, Brno, Czech Republic, pp. 23–30 (2004)
17. Zerbitzuak, E.H. (ed.): Elhuyar Zientzia eta Teknologiaren Hiztegi Entziklopedikoa. Elhuyar Edizioak/Euskal Herriko Unibertsitatea (2009)
18. Agirre, E., Soroa, A.: Personalizing PageRank for Word Sense Disambiguation. In: Proceedings of EACL 2009, Athens, Greece, pp. 33–41 (2009)