Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation

Ruslan Mitkov, Le An Ha, Andrea Varga and **Luz Rello**
University of Wolverhampton
Wolverhampton, UK
{R.Miktov, L.A.Ha, Andrea.Varga,
L.RelloSanchez}@wlv.ac.uk

Abstract

Mitkov and Ha (2003) and Mitkov et al. (2006) offered an alternative to the lengthy and demanding activity of developing multiple-choice test items by proposing an NLP-based methodology for construction of test items from instructive texts such as textbook chapters and encyclopaedia entries. One of the interesting research questions which emerged during these projects was how better quality distractors could automatically be chosen. This paper reports the results of a study seeking to establish which similarity measures generate better quality distractors of multiple-choice tests. Similarity measures employed in the procedure of selection of distractors are collocation patterns, four different methods of WordNet-based semantic similarity (extended gloss overlap measure, Leacock and Chodorow's, Jiang and Conrath's as well as Lin's measures), distributional similarity, phonetic similarity as well as a mixed strategy combining the aforementioned measures. The evaluation results show that the methods based on Lin's measure and on the mixed strategy outperform the rest, albeit not in a statistically significant fashion.

1 Introduction

Multiple-choice tests are sets of test items, the latter consisting of a question or *stem* (e.g. Who was voted the best international footballer for 2008?), the correct *answer* (e.g. Ronaldo) and *distractors* (e.g. Messi, Ronaldino, Torres). This type of test has proved to be an efficient tool for measuring students' achievement and is used on a daily basis both for assessment and diagnostics worldwide.[1] According to Question Mark Computing Ltd (p.c.), who have licensed their Perception software to approximately three million users so far, 95% of their users employ this software to administrate multiple-choice tests.[2] Despite their popularity, the manual construction of such tests remains a time-consuming and labour-intensive task. One of the main challenges in constructing a multiple-choice test item is the selection of plausible alternatives to the correct answer which will better distinguish confident students from unconfident ones.

Mitkov and Ha (2003) and Mitkov et al. (2006) offered an alternative to the lengthy and demanding activity of developing multiple-choice test items by proposing an NLP-based methodology for construction of test items from instructive texts such as textbook chapters and encyclopaedia entries. This methodology makes use of NLP techniques including shallow parsing, term extraction, sentence transformation and semantic distance computing and employs resources such as corpora and ontologies like WordNet. More specifically, the system identifies important terms in a textbook text,

[1] This paper is not concerned with the issue of whether multiple-choice tests are better assessment methodology that other types of tests. What it focuses is on improving our new NLP methodology to generate multiple-choice tests about facts explicitly stated in single declarative sentences by establishing which semantic similarity measures give rise to better distractors.

[2] More information on the Perception software can be found at: www.questionmark.com/perception

Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, pages 49–56,
Athens, Greece, 31 March 2009.

transforms declarative sentences into questions and mines for terms which are semantically close to the correct answer, to serve as distractors.

The system for generation of multiple-choice tests described in Mitkov and Ha (2003) and in Mitkov et al. (2006) was evaluated in practical environment where the user was offered the option to post-edit and in general to accept, or reject the test items generated by the system[3]. The formal evaluation showed that even though a significant part of the generated test items had to be discarded, and that the majority of the items classed as 'usable' had to be revised and improved by humans, the quality of the items generated and proposed by the system was not inferior to the tests authored by humans, were more diverse in terms of topics and very importantly – their production needed 4 times less time than the manually written items. The evaluation was conducted both in terms of measuring the time needed to develop test items and in terms of classical test analysis to assess the quality of test items.

The paper is structured as follows. Section 2 will outline the importance of distractors in multiple-choice testing as the different strategies for automatic selection of the distractors are the subject of this study. Section 3 will describe how test items are produced and will detail the different strategies (semantic similarity measures and phonetic similarity) used for the selection of distractors. Section 4 outlines the in-class experiments, presents the evaluation methodology, reports on the results and discusses these results.

2 The importance of quality distractors

One of the interesting research questions which emerged during the above research was how better quality distractors could automatically be chosen. In fact user evaluation showed that from the three main tasks performed in the generation of multiple-choice tests (term identification, sentence transformation and distractor selection), it was distractor selection which needed further improvement with a view to putting it in practical use.

Distractors play a vital role for the process of multiple-choice testing in that good quality distractors ensure that the outcome of the tests provides more credible and objective picture of the knowledge of the testees involved. On the other hand, poor distractors would not contribute much to the accuracy of the assessment as obvious or too easy distractors will pose no challenge to the students and as a result, will not be able to distinguish high performing from low performing learners.

The principle according to which the distractors were chosen, was semantic similarity (Mitkov and Ha, 2003). The semantically closer were the distractors to the correct answer, the most 'plausible' they were deemed to be. The rationale behind this consists in the fact that distractors semantically distant from the correct answer could make guessing a 'straightforward task'. By way an example, if processing the sentence 'Syntax is the branch of linguistics which studies the way words are put together into sentences', the multiple-choice generation system would identify *syntax* as an important term, would transform the sentence into the question 'Which branch of linguistics studies the way words are put together into sentences?' and would choose 'Pragmatics', 'Morphology' and 'Semantics' as distractors to the correct answer 'Syntax', being closer to it than 'Chemistry', 'Football' or 'Beer' for instance (which if offered as distractors, would be easily dismissed by people who do not have even any knowledge of linguistics).

While the semantic similarity premise appears as a logical way forward to automatically select distractors, there are different methods or measures which compute semantic similarity. Each of these methods could be evaluated individually but here we evaluate their suitability for the task of selection of distractors in multiple-choice tests. This type of evaluation could be regarded as *extrinsic evaluation* of each of the methods, where the benchmark for their performance would not be an annotated corpus or human judgement on accuracy, but to what extent a specific NLP application can benefit from employing a method.

Another premise that this study seeks to verify is whether orthographically close distractors, in addition to being semantically related, could yield even better results.

[3] A post-editor's interface was developed to this end.

3 Production of test items and selection of distractors

Test items were constructed by a program based on the methodology described in the previous section. We ran the program on an on-line course materials in linguistics (Vajda, 2001). A total of 144 items were initially generated. 31 out of these 144 items were kept for further considerations as they either did not need any or, only minor revision. The remaining 113 items were deemed to require major post-editing revision. The 31 items kept for consideration were further revised by a second linguist and finally, we narrowed down the selection to 20 questions for the experiments[4]. These 20 questions gave a rise to a total of eight different assessments. Each assessment had the same 20 questions but they differed in the sets of distractors as these were chosen using different similarity measures[5] (sections 3.1-3.5).

To generate a list of distractors for single-word terms the function *coordinate terms* in WordNet is employed. For multi-word terms, noun phrases with the same head as the correct answers appearing in the source text as well as entry terms from Wikipedia having the same head with the correct answers, are used to compile the list of distractors. This list of distractors is offered to the user from which he or she could choose his/her preferred distractors.

In this study we explore which is the best way to narrow down the distractors to the 4 most suitable ones. To this end, the following strategies for computing semantic (and in one case, phonetic) similarity were employed: (i) collocation patterns, (ii-v) four different methods of WordNet-based semantic similarity (Extended gloss overlap measure, Leacock and Chodorow's, Jiang and Conrath's and Lin's measures), (vi) Distributional Similarity, and (vii) Phonetic similarity.

[4] The following is an example of an item generated of the program and then post-edited.
"Which type of clause might contain verb and dependent words? i) verb clause ii) adverb clause iii) adverbial clause
iv) multiple subordinate clause v) subordinate clause".

[5] It should be noted that there were cases where the different selection/similarity strategies picked the same distractors.

3.1 Collocation patterns

The collocation extraction strategy used in this experiment is based on the method reported in (Mitkov and Ha, 2003). Distractors that appear in the source text are given preference. If there are not enough distractors, distractors are selected randomly from the list.

For the other methods described below (sections 3.2-3.5), instead of giving preference to noun phrases appearing in the same text, and randomly pick the rest from the list, we ranked the distractors in the list based on the similarity scores between each distractor and the correct answer and chose the top 4 distractors.

We compute similarity for words rather than multi-word terms. When the correct answers and distractors are multi-word terms, we calculate the similarities between their modifier words. By way of example, in the case of "verb clause" and "adverbial clause", the similarity score between "verb" and "adverbial" is computed. When the correct answer or distractor contains more than one modifiers we compute the similarity for each modifier pairs and we choose the maximum score. (e.g. for "verb clause" and "multiple subordinate clause", similarity scores of "verb" and "multiple" and of "verb" and "subordinate" are calculated, the higher one is considered to represent the similarity score).

3.2 Four different methods for WordNet-based similarity

For computing WordNet-based semantic similarity we employed the package made available by Ted Pedersen[6]. Pedersen's tool computes (i) extended gloss overlap measure (Banerjee and Pedersen, 2003), (ii) Leacock and Chodorow's (1998) measure, (iii) Jiang and Conrath's (1997) measure and (iv) Lin's (1997) measure.

The extended gloss overlap measure calculates the overlaps between not only the definitions of the two concepts measured but also among those concepts to which they are related. The relatedness score is the sum of the squares of the overlap lengths.

Leacock and Chodorow's measure uses the normalised path length between the two concepts c_1 and c_2 and is computed as follows:

[6] http://search.cpan.org/~tpederse/WordNet-Similarity

$$sim_{lch}(c_1, c_2) = -\log\left[\frac{len(c_1, c_2)}{(2 \times MAX)}\right] \quad (1)$$

where *len* is the number of edges on the shortest path in the taxonomy between the two concepts and *MAX* is the depth of the taxonomy.

Jiang and Conrath's measure compares the sum of the information content of the individual concepts with that of their lowest common subsumer:

$$sim_{jcn}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)} \quad (2)$$

where *IC(c)* is the information content (Patwardhan et al., 2003) of the concept *c*, and *lcs* denotes the lowest common subsumer, which represents the most specific concept that the two concepts have in common.

The Lin measure scales the information content of lowest common subsumer with the sum of information content of two concepts.

$$sim_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (3)$$

3.3 Distributional similarity

For computing distributional similarity we made use of Viktor Pekar's implementation[7] based on Information Radius, which according to a comparative study by Dagan et al. (1997) performs consistently better than the other similar measures. Information Radius (or Jensen-Shannon divergence) is a variant of Kullback-Leiber divergence measuring similarity between two words as the amount of information contained in the difference between the two corresponding co-occurrence vectors. Every word w_j is presented by the set of words $w_{i1...n}$ with which it co-occurs. The semantics of w_j are modelled as a vector in an *n*-dimensional space where *n* is the number of words co-occurring with w_j, and the features of the vector are the probabilities of the co-occurrences established from their observed frequencies, as in (4). In Pekar's implementation, if one word is identified as dependent on another word by a dependency parser, these two words are said to be "co-occuring"[8]. The corpus used to collect the co-occurance vector was the BNC and the dependency parsed used the FDG parser (Tapanainen and Järvinen, 1997). The Information Radius (JS) is calculated using (5).

$$C(w_j) = \left\langle P(w_j \mid w_{i1}), P(w_j \mid w_{i2}), ..., P(w_j \mid w_{in}) \right\rangle \quad (4)$$

$$JSD(C(w_j) \,||\, C(w_k)) = \frac{1}{2} D(C(w_j) \,||\, M) + \frac{1}{2} D(C(w_k) \,||\, M) \quad (5)$$

where $M = \frac{1}{2}(C(w_j) + C(w_k))$

3.4 Phonetic similarity

For measuring phonetic similarity we use Soundex, phonetic algorithm for indexing words by sound. It operates on the principle of term based evaluation where each term is given a Soundex code. Each Soundex code itself consists of a letter and three numbers between 0 and 6. By way of example the Soundex code of *verb* is *V610* (the first character in the code is always the first letter of the word encoded). Vowels are not used and digits are based on the consonants as illustrate by the following table:

1. B, P, F, V
2. C, S, K, G, J, Q, X, Z
3. D, T
4. L
5. M, N
6. R

Table 1 Digits based on consonants

First the Soundex code for each word is generated[9]. Then similarity is computed using the Difference method, returning an integer result ranging in value from 1 (least similar) to 4 (most similar).

3.5 Mixed Strategy

After items have been generated by the above seven methods, we pick three items from each method, except from Soundex, where only two items have been picked, to compose an

[7] http://clg.wlv.ac.uk/demos/similarity/index.html

[8] There are many other ways to construct the co-occurrence vectors. This paper does not intend to exploit these different ways.

[9] We adopt the phonetic representation used in MS SQL Server. As illustrated above, each soundex code consists of a letter and three numbers, such as A252.

assessment of 20 items. This assessment is called "mixed", and used to assess whether or not an assessment with distractors generated by combining different methods would produce a different result from an assessment featuring distractors generated by a single method.

4 In-class experiments, evaluation, results and discussion

The tests (papers) generated with the help of our program with the distractors chosen according the different methods described above, were taken by a total of 243 students from different European universities: University of Wolverhampton (United Kingdom), University College Ghent (Belgium), University of Saarbrücken (Germany), University of Cordoba (Spain), University of Sofia (Bulgaria). A prerequisite for the students taking the test was that they studied language and linguistics and that they had a good command of English. Each test paper consisted of 20 questions and the students had 30 minutes to reply to the questions. The tests were offered through the Questionmark Perception web-based testing software which in addition to providing a user-friendly interface, computes diverse statistics related to the test questions answered.

In order to evaluate the quality of the multiple-choice test items generated by the program (and subsequently post-edited by humans), we employed standard *item analysis*. Item analysis is an important procedure in classical test theory which provides information as to how well each item has functioned. The item analysis for multiple-choice tests usually consists of the following information (Gronlund, 1982): (i) the difficulty of the item, (ii) the discriminating power and (iii) the usefulness[10] of each distractor. This information can tell us if a specific test item was too easy or too hard, how well it discriminated between high and low scorers on the test and whether all of the alternatives functioned as intended. Such types of analysis help improve test items or discard defective items.

Whilst this study focuses on the quality of the distractors generated, we believe that the distractors are essential for the quality of the overall test and hence the *difficulty* of an item and its *discriminating power* are deemed appropriate to assess the quality of distractors, even though the quality of the test stem also pays in important part. On the other hand usefulness is a completely independent measure as it looks at distractors only and not only the combination of stems and distractors.

In order to conduct this type of analysis, we used a simplified procedure, described in (Gronlund, 1982). We arranged the test papers in order from the highest score to the lowest score. We selected one third of the papers and called this the upper group. We also selected the same number of papers with the lowest scores and called this the lower group. For each item, we counted the number of students in the upper group who selected each alternative; we made the same count for the lower group.

(i) Item Difficulty

We estimated the *Item Difficulty* (ID) by establishing the ratio of students from the two groups who answered the item correctly (ID = C/T, where C is the number who answered the item correctly and T is the total number of students who attempted the item). As Table 2 shows, from the items featuring distractors generated using the collocation method[11], there were 4 too easy and 0 too difficult items.[12] The average Item Difficulty was 0.61. From the items with distractors generated using WordNet-based similarity[13], the results were the following. When employing the extended gloss overlap measure there were 2 too easy and 0 too difficult items, showing an average ID of 0.58. Leacock and Chodorow's measure produced 1 too easy and 3 too difficult items with item average difficulty of 0.54. The use of Jiang and Conrath's measure resulted in 3 too easy and 1 too difficult items; the average item difficulty observed was 0.57. Lin's measure delivered the best results from the

[10] Originally called 'effectiveness'. We chose to term this type of analysis 'usefulness' to distinguish it from the (cost/time) 'effectiveness' of the semi-automatic procedure as opposed to the manual construction of tests.

[11] Henceforth referred to as 'collocation items'; the distractors generated are referred to as 'collocation distractors'.

[12] For experimental purposes, we consider an item to be 'too difficult' if ID ≤ 0.15 and an item 'too easy' if ID ≥ 0.85.

[13] Henceforth referred to as 'WordNet items'; the distractors are referred to as 'WordNet distractors'.

point of item difficulty with an almost ideal average item difficulty of 0.51 (the recommended item difficult is 0.5; see also footnote 16); there were 2 too easy and 1 too difficult items.

The items constructed on the basis of distractors selected via the distributional similarity metric[14], scored an average ID of 0.64 with 6 items being too easy and 1 — too difficult. From the items with distractors produced using the phonetic similarity algorithm[15], there were 4 too easy and 0 too difficult questions with overall average difficult of 0.60. Finally, a mixed strategy produced test items with average difficulty of 0.53, 1 of them being too easy and 0 — too difficult.

The results showed that almost all items produced after selecting distractors using the strategies described above, featured very reasonable ID values. In many cases the average values were close to the recommended ID value of 0.5 with Lin's measure delivering the best ID of 0.51. Runners-up are the mixed strategy delivering items with average ID 0.53 Leacock and Chodorow's measure contributing to the generation of items with average ID of 0.54.

(ii) Discriminating Power

We estimated the item's *Discriminating Power* (DP) by comparing the number students in the upper and lower groups who answered the item correctly. It is desirable that the discrimination is *positive* which means that the item differentiates between students in the same way that the total test score does.[16] The formula for computing the *Discriminating Power* is as follows: DP = $(C_U - C_L)$: T/2, where C_U is the number of students in the upper group who answered the item correctly and C_L the number of the students in the lower group that did so. Here again T is the total number of students included in the item analysis.[17]

The average Discriminating Power for the collocation items was 0.33 and there were no negative discriminating collocation test items.[18] The figures associated to the WordNet items were as follows. The average DP for items produced with the extended gloss overlap measure was 0.32, and there were 2 items with negative discrimination. Leacock and Chodorow's measure did not produce any items with negative discrimination and the average DP of these was 0.38. Jiang and Conrath's measure gave rise to 2 negatively discriminating items and the average DP of the items based on this measure was 0.29. The selection of distractors with Lin's measure resulted in items with average DP of 0.37; none of them had a negative discrimination.

The average discrimination power for the distributional items was 0.29 (1 item with negative discrimination) and for phonetic items – 0.34 (0 item with negative discrimination). The employment of mixed strategy when selecting distractors which resulted in items with average DP of 0.39 (0 items with negative discrimination).

The figures related to the Discriminating Power of the items generated showed that whereas the DP was not of the desired high level, as a whole the proportion of items with negative discrimination was fairly low (Table 2). The items did not differ substantially in terms of the values of DP, the top performer being the items where the distractors were selected on the basis of the mixed strategy, followed by those selected by Leacock and Chodorow's measure and phonetic similarity.

(iii) Usefulness of the distractors

The *usefulness of the distractors* is estimated by comparing the number of students in the upper and lower groups who selected each incorrect alternative. A good distractor should attract more students from the lower group than the upper group.

The evaluation of the distractors estimated the average difference between students in the

[14] Henceforth referred to as 'distributional items'; the distractors are referred to as 'distributional distractors'.

[15] Henceforth referred to as 'phonetic items'; the distractors are referred to as 'phonetic distractors'.

[16] Zero DP is obtained when an equal number of students in each group respond to the item correctly. On the other hand, negative DP is obtained when more students in the lower group than the upper group answer correctly. Items with zero or negative DP should be either discarded or improved.

[17] Maximum positive DP is obtained only when all students in the upper group answer correctly and no one in the lower group does. An item that has a maximum DP (1.0) would have an ID 0.5; therefore, test authors are advised to construct items at the 0.5 level of difficulty.

[18] Obviously a negative discriminating test item is not regarded as a good one.

	Item Difficulty			Item Discriminating Power		Usefulness of distractors		
	average item difficulty	too easy	too difficult	average discriminating power	negative discriminating power	poor	not useful	average difference
Collocation items	0.61	4	0	0.33	0	2	24	0.74
WordNet items								
- Extended gloss overlap	0.58	2	0	0.32	2	9	17	0.71
- Leacock and Chodorow	0.54	1	3	0.38	0	9	20	0.76
- Jiang and Conrath	0.57	3	1	0.29	2	10	19	0.71
- Lin	0.51	2	1	0.37	0	10	16	0.83
Distributional items	0.64	6	1	0.29	1	6	27	0.79
Phonetic items	0.60	4	0	0.34	0	5	31	0.66
Mixed strategy items	0.53	1	0	0.39	0	5	14	0.89

Table 2: Item analysis

lower and upper groups to be 0.74 for the sets of distractors generated using collocations. For the WordNet distractors the results were as follows. The average distance between the students in the lower and upper groups was found to be 0.71 for the extended gloss overlap distractors, 0.76 for the Leacock and Chodorow distractors, 0.71 for the Jiang and Conrath distractors and 0.83 for the Lin distractors. For the distractors selected by way of distributional similarity the average difference between students in the lower and upper groups was 0.79, for the phonetic distractors — 0.66 and for those selected by a mixed strategy — 0.89.

In our evaluation we also used the notions of *poor distractors* as well as *not-useful* distractors. Distractors are classed as *poor* if they attract more students from the upper group than from the lower group. There were 2 (2.5%) poor distractors from the collocation distractors. The WordNet distractors fared as follows with regard to the number of poor distractors. There were altogether 9 (11%) poor distractors from the extended gloss overlap distractors, 9 (11%) from the Leacock and Chodorow distractors, 10 (12%) from the Jiang and Conrath distractors and 10 (12%) from the Lin ones. There were 6 (7.5%) from the distributional similarity which were classed as poor, 5 (6%) from the phonetic similarity ones were classed as poor and 5 (6%) from the distractors selected through a mixed strategy were classed as such (Table 2).

On the other hand, distractors are termed *not useful* if they are not selected by any students at all. The evaluation showed (see Table 2) that there were 24 (30%) distractors deemed not useful from the collocation distractors. The figures for not useful distractors for those selected by way of WordNet similarity were as follows: 17 (21%) for extended gloss overlap distractors, 20 (25%) for the Leacock and Chodorow distractors, 19 (24%) for the Jiang and Conrath distractors and 16 (20%) for the Lin ones. From the distributional distractors, 27 (34%) emerged as not useful, whereas 31 (39%) phonetic similarity and 14 (18%) mixed strategy distractors were found not useful.

The overall figures suggest that the 'most useful' distractors are those chosen with mixed strategy (highest average difference 0.89; lowest number of not useful distractors, second lowest number of poor distractors), followed by those chosen with Lin's WordNet measure (second highest average distance of 0.83; second lowest number of not useful distractors).

Summarising the results of the item analysis, it is clear that there is not a method that outperforms the rest in terms of producing best quality items or distractors. At the same time it is also clear that in general the mixed strategy and Lin's measure consistently perform better than the rest of methods/measures. Phonetic similarity did not deliver as expected.

Although the results indicate that the Lin items have the best average item difficulty, none of the difference (between item difficulty of Lin and other methods, or between any pair of methods) is statistically significant. From the DP point of view, only the difference between mixed strategy (0.39) and distributional items (0.29) is statistically significant ($p<0.05$). For the distractor usefulness measure, none of the difference is statistically significant ($p<0.05$).

5 Conclusion

In this study we conducted extrinsic evaluation of several similarity methods (collocation patterns; four different methods of WordNet-based semantic similarity: extended gloss overlap measure, Leacock and Chodorow's, Jiang and Conrath's as well as Lin's measures; distributional similarity; phonetic similarity; mixed strategy) by seeking to establish which one would be most suitable for the task of selection of distractors in multiple-choice tests. The evaluation results based on item analysis suggests that whereas there is not a method that clearly outperforms in terms of delivering better quality distractors, mixed strategy and Lin's measure consistently perform better than the rest of methods/measures. However, these two methods do not offer any statistically significant improvement over their closest competitors.

Acknowledgments

We would like to express our gratitude to Kathelijne Denturck, Johann Haller, Veronique Hoste, Constantin Orasan, Miriam Seghiri, Andrea Stockero and Irina Temnikova for helping us in the organisation of the in-class experiments.

References

Banerjee, S. and Perderson, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 805-810.

Dagan I., Lee L., and Pereira F. 1997. Similarity-based methods for word sense disambiguation. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain, 56-63.

Gronlund, N. 1982. *Constructing achievement tests*. New York: Prentice-Hall Inc.

Jiang, J. and Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the International Conference on Research in Computational Linguistics. Taiwan*, 19-33.

Lin, D. 1997. Using syntactic dependency as a local context to resolve word sense ambiguity. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain, 4-71.

Leacock, C., Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C., 1998, *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 265–283.

Mitkov R. and Ha L.A. 2003. Computer-aided generation of multiple-choice tests. *Proceedings of the HLT/NAACL 2003 Workshop on Building educational applications using Natural Language Processing*. Edmonton, Canada, 17-22.

Mitkov, R., An, L.A. and Karamanis, N. 2006. "A computer-aided environment for generating multiple-choice test items". *Journal of Natural Language Engineering,* 12 (2): 177-194.

Patwardhan, S, Banerjee, S. and Pedersen, T. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics.* Mexico City, 241-257.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence.* Montreal, 448–453.

Tapanainen, P. and Järvinen, T. 1997. A non-projective dependency parser. *Proceedings of the 5th Conference of Applied Natural Language Processing,* Washington, 64-71.

Vajda, E.J. 2001 Course Materials from the module of Introduction to Linguistics. Professor Edward J. Vajda Homepage, Washington, Western Washington University, Modern and Classical Languages.http://pandora.cii.wwu.edu/vajda/ling201/ling201home.htm