Paraphrase Extraction from Validated Question Answering Corpora in Spanish*

Jesús Herrera, Anselmo Peñas, Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
C/ Juan del Rosal, 16, E-28040 Madrid
{jesus.herrera, anselmo, felisa}@lsi.uned.es

Resumen: Partiendo del debate sobre la definición de paráfrasis, este trabajo intenta clarificar lo que las personas consideran como paráfrasis. El experimento realizado parte de una de las distintas campañas que generan cada año grandes cantidades de datos validados, susceptibles de ser reutilizados con diferentes fines. En este artículo se describe con detalle un método simple –fundamentado en reconocimiento de patrones y operaciones de inserción y eliminación–, capaz de extraer una importante cantidad de paráfrasis de corpora de Pregunta–Respuesta evaluados. Se muestra además la evaluación realizada por expertos del corpus obtenido. Este trabajo ha sido realizado para el español.
Palabras clave: Extracción de paráfrasis, corpus de Pregunta–Respuesta, definición de paráfrasis

Abstract: Basing on the debate around the definition of paraphrase, this work aims to empirically clarify what is considered a paraphrase by humans. The experiment accomplished has its starting point in one of the several campaigns that every year generate large amounts of validated textual data, which can be reused for different purposes. This paper describes in detail a simple method –based on pattern–matching and deletion and insertion operations–, able to extract a remarkable amount of paraphrases from Question Answering assessed corpora. An assessment of the corpus obtained was accomplished by experts, and an analysis of this process is shown. This work has been developed for Spanish.
Keywords: Paraphrase extraction, Question Answering corpus, paraphrase definition

1 Introduction

The main idea of the present work is that, although several definitions of the concept of paraphrase have been already made, it is still important to determine what humans understand when they are said to evaluate if a pair of statements are related by a paraphrase relationship. For this purpose, it was decided to obtain a corpus containing pairs of statements that could be paraphrases; these pairs were be assessed by experts in order to determine if, effectively, there was a paraphrase relationship between them. In addition, it was considered that some corpora could successfully be reused in order to automatically extract these pairs of candidates for paraphrase. The corpus ed was the corpus of assessed answers –in Spanish– from the Question Answering (QA) exercise proposed in the 2006 edition of the Cross Language Evaluation Forum (CLEF). The experiment accomplished suggests that with such corpus it is viable to obtain a high amount of paraphrases with a fully automated and simple process. Only shallow techniques were applied all along this work for this first approach. This method increases the set of proposals for paraphrase obtention given until now, for example: (Barzilay and McKeown, 2001) and (Pang et al., 2003) used text alignment in different ways to obtain paraphrases; (Lin and Pantel, 2001) used mutual information of word distribution to calculate the similarity of expressions,

* We are very grateful to Sadi Amro Rodríguez, Mónica Durán Mañas and Rosa García–Gasco Villarrubia for their contribution by assessing the paraphrase corpus. We also would like to thank Claudia Toda Castán for revising this text. This work has been partially supported by the Spanish Ministry of Science and Technology within the project R2D2–SyEMBRA (TIC–2003–07158–C04–02), and by the Regional Government of Madrid under the auspices of MAVIR Research Network (S–0505/TIC–0267).

(Ravichandran and Hovy, 2002) used pairs of questions and answers to obtain varied patterns which give the same answer; and (Shinyama et al., 2002) obtained paraphrases by means of named entities found in different news articles reporting the same event.

In section 2 an overview of the experiment is given. Section 3 describes all the steps accomplished in order to transform the multilingual source corpus in a monolingual corpus of paraphrase candidates, ready to be assessed. Section 4 describes the activity developed by the assessors and the results obtained; the problems detected in the process are listed, with suggestions for its improvement; and, finally, some ideas about what humans understand under the concept of paraphrase are outlined. In section 5 some conclusions and proposals for future work are given.

2 The experiment

Every year, QA campaigns like the ones of the CLEF (Magnini et al., 2006), the Text REtrieval Conference (TREC) (Voorhees and Dang, 2005) or the NII–NACSIS Test Collection for IR Systems (NTCIR) (Fukumoto et al., 2004) (Kato et al., 2004), generate a large amount of human–assessed textual corpora. These corpora, containing validated information, can be reused in order to obtain data that can be well-spent by a wide range of systems. The idea, given by (Shinyama et al., 2002), that articles derived from different newspapers can contain paraphrases if they report the same event, made us aware of the fact that in the QA campaign of the CLEF the participating systems usually obtain several answers for a certain question; the answers, taken from a news corpus, are related by the common theme stated by this question. Thus, probably a remarkable number of these answers will compose one or more sets of paraphrases. But, is it easy for a computer program to extract that information? This last question motivated a study of the corpora available after the assessments of the Question Answering exercise of the CLEF (QA@CLEF) 2006 campaign. The first action accomplished aimed at determine if, by means of simple techniques, a corpus of candidates for paraphrases could be obtained in a fully automatic way. After it, this corpus was evaluated by three philologists in order to detect the exact set of paraphrases obtained, i.e., the candidates that were, efectively, paraphrases; their judgements were used as a voting to obtain this final set. The output of this assessment process was used to try to identify what humans understand under "paraphrase".

3 Building a corpus for the experiment

One of the objectives of the experiment was to determine the best way to obtain a paraphrase corpus from a QA assessed corpus using shallow techniques. It was accomplished as described in the following subsections.

3.1 The multilingual source corpus

The assessment process of the QA@CLEF produces a multilingual corpus with its results. This QA corpus contains, for every language involved in the exercise, the following data: the questions proposed, all the answers given to every question, and the human assessment given to every answer (right, wrong, unsupported, inexact) (Magnini et al., 2006). Our idea was to use this corpus as a source to obtain a paraphrase corpus in Spanish.

3.2 The Spanish corpus

Since the QA@CLEF is a multiple language campaign and the scope of our experiment covered only the Spanish language, we extracted from the source corpus all the questions and assessed answers in Spanish. Thus, a monolingual Spanish corpus –which is a subcorpus of the source one– was ready to be used. The assessed answers were represented in the format shown in figure 1; for every answer there is a record in the file consisting of the following fields, from left to right and separated by tab blanks: the calification given by a human assessor, the number of the question, the identification of the run and the system, the confidence value, the identification of the document that supports the answer, the answer and the snippet from the indicated document that contains the given answer.

This format follows the one established for the QA@CLEF 2006[1].

3.3 Extraction of validated data

The first action over the Spanish corpus was to select the records containing at least one answer assessed as correct. Thus, only

[1]Guidelines of QA@CLEF 2006: `http://clefqa.itc.it/guidelines.html`

Figure 1: Excerpt of the Spanish corpus.

```
...
R   0065   inao061eses   1.00   EFE19940520−12031   moneda griega
...GRECIA−MONEDA INTERVENCION BANCO CENTRAL PARA SALVAR DRACMA Atenas
, 20 may (EFE).− El Banco de Grecia (emisor) tuvo que intervernir hoy
, viernes , en el mercado cambiario e inyectar 800 millones de marcos
alemanes para mantener el valor del dracma , moneda griega , tras
la liberación de los movimientos del capital el pasado lunes ....
...
```

human–validated data were considered for the experiment. From the 200 questions proposed to the systems participating in the QA@CLEF 2006, 153 obtained one or more correct answers by one or more systems. From every selected record, the answer and the snippet containing it were extracted, because all the textual information liable to contain paraphrases is included into them.

3.4 Data transformation and selection

After it, every answer was turned into its affirmative version by means of very simple techniques, following the initial idea of high simplicity for this work. First of all, punctuation signs were deleted. The most frequent ones were *¿* and *?*. Next, a list of frecuencies of interrogative formulations in Spanish was made in order to establish a set of rules for turning them into the affirmative form. Two transformation operations were applied by means of these rules: deletion and insertion. These operations affect only to the initial words of the questions. Thus, for example, if the first words of a question are *"quién es"*, they must just be deleted for obtaining the affirmative version; but, if the first words of a question are *"qué"* + substantive + verb, the word *"qué"* must be deleted and the word *"que"* must be inserted after the substantive and before the verb. Thus, once deleted the punctuation signs and applied the previous rule to the question *¿qué organización dirige Yaser Arafat?* (what organization leads Yasser Arafat?), its affirmative form is as follows: *organización que dirige Yaser Arafat* (organization leaded by Yasser Arafat). Some rules are very easy to obtain, such as the previous one, but some others are quite difficult; for example, when a question starts with the word *cuándo* (when), it is not trivial to transform it into an affirmative form, because several options exist and it is not possible to decide what is the more appropriate without a semantic analysis. The question *¿cuándo murió Stalin?* (when did Stalin dead?) serves to illustrate this situation; it could be transformed into different affirmative forms: *fecha en la que murió Stalin* (date in which Stalin die), *momento en el que murió Stalin* (moment in which Stalin died), etcetera. Thus, it was decided to apply the following rule: if a question starts with the word *cuándo*, then delete *cuándo*; therefore, for the present example, the question *¿cuándo murió Stalin?* is transformed into *murió Stalin* (Stalin died). This was considered the best approach that could be obtained using only surface techniques. Some of the 29 rules identified are shown in table 1. This list of rules raises from a research work over the Spanish corpus described, and more rules could be identified in future related works with other corpora.

Once applied the previous rules over the corpus, it was identified a set of monograms and bigrams that must be deleted when appearing at the beginning of the new statements obtained. The monograms are articles (*"el"*, *"la"*, *"lo"*, *"los"*, *"las"*), and the bigrams are combinations of the verb *"ser"* (to be) followed of an article, for example: *"era el"*, *"es la"*, *"fue el"*. Thus, for example, once deleted the punctuation signs, the application of rule number 1 from table 1 to the question *¿qué es el tóner?* (what is toner?), we obtained the following statement: *el tóner* (the toner); then, the article *"el"* is deleted and the definitive statement is *tóner* (toner).

Since the techniques used for turning the questions into their affirmative form were only at the lexical level, slightly agrammatical statemens were produced. Anyway, most of the errors consist of a missing article or relative pronoun. Nevertheless, a human can perfectly understand this kind of agrammatical statements and, in addition, a lot of sys-

Table 1: Some rules identified for automatic conversion into the affirmative form.

#	**If the first words of the question are:**	**Then:**
1	*qué es*	delete *qué es*
2	*qué* + substantive + verb	delete *qué* insert *que* after the substantive and before the verb
3	*a qué* + substantive + verb	delete *a qué* insert a *que* after the substantive and before the verb
4	*quién es*	delete *quién es*
5	*cuántos* + list of words + verb	delete *cuántos* insert *número de* at the beginning insert *que* after the list of words and before the verb
6	*cuándo*	delete *cuándo*
7	*nombre*	delete *nombre*
8	*dé*	delete *dé*

tems do not consider stopwords (where articles and/or relative pronouns are usually included). These errors can be avoided applying a morphological analysis; but we preserved them, appart from for the sake of simplicity, in order to permit a future study of the importance of their presence in the corpus. For example: can systems using the corpus accomplish their tasks despite the presence of some grammatical errors in it? If so, the morphological analysis could be avoided for building such kind of corpora. At this point an interesting suggestion arises: campaigns such the Answer Validation Exercise (AVE) (Peñas et al., 2006), developed for the first time within the 2006 CLEF, need an important human effort for transforming the answers from the associated QA exercise into their affirmative form. Therefore, the method implemented for this experiment could e a useful tool for tasks such the AVE.

After turning the questions into there affirmative form, a normalization and filter action was accomplished over the corpus in order to avoid the frequent phenomenon of having a set of equal –or very similar– answers given by different systems to a determined question. It consisted of the following steps:

1. Lowercase the affirmative version of all the questions, and all the answers.
2. Eliminate punctuation signs and particles such as articles or prepositions at the beginning and the end of every statement.
3. For the set of normalized answers associated to every question, eliminate the repeated ones and the ones contained by other. That is, if the string representing the answer is the same or is a substring of other string representing the answer and pertaining to the set of answers for a determined question, the former one is eliminated from the set of answers.

After the normalization and filtering, a first inspection of the corpus obtained was accomplished in order to determine if more operations should be done for obtaining paraphrases. At the beginning it may seem that little work is to be done with the questions in affirmative form and the answers. But previous works on paraphrase detection suggested that the longest common subsequence of a pair of sentences could be considered for the objectives of this work (Bosma and Callison–Burgh, 2006) (Zhang and Patrick, 2005). A first set of tests using the longest common subsequence showed that some anwers could be exploited to augment the amount of paraphrases; for example, *presidente de Brasil* (president of Brazil) is a reformulation for *presidente brasileño* (Brazilian president) and, if the largest common subsequence is deleted from both statements, *de Brasil* (of Brazil) and *brasileño* (Brazilian) are the new statements obtained, and they are a paraphrase of each other. The problem is that it is necessary to determine what statements are good candidates for such operation, and it is not easy by using simple techniques. In addition, little examples of this kind were found; thus, no much information could be added. This is because this operation was not considered for the present work.

3.5 What does not work?

The previous idea about deleting the largest common subsequence from a pair of strings

in order to find paraphrases made arise the following intuition: when two texts contain the same information, if the common words are deleted, the rest of the words conform a pair of strings that could –perhaps– be a pair of paraphrases. The snippets of the corpus were tested to determine if such intuition was correct. The test consisted of grouping all the snippets related to every question and, then, taking every possible pair of snippets among the ones pertaining to the same group, deleting the largest common subsequence of the pair. An examination of the output of this operation revealed that it was improductive to obtain paraphrases. At this point the value for the present work of the previous labour accomplished by the QA systems becomes patently clear, because they filter information from the snippets and virtually there is no need to treat it "again". Therefore it was decided not to use the snippets for the paraphrase searching, but only the questions into its affirmative form and the different given answers.

3.6 The final corpus

After applying the operations described in subsection 3.4 over the validated data from the Spanish subcorpus, the definitive corpus for this work was ready. It consisted of groups of related statemens; each group contained the affirmative form of a question and all the different answers obtained from the participating systems. Giving some numbers, this corpus shows 87 groups of statemes for which 1 answer was given to the question, 47 groups with 2 different answers for the question, 12 groups with 3 answers, 5 groups with 4 answers, 1 group with 1 answer, no groups with 6 answers and 1 group with 7 answers. None of the considered questions (see subsection 3.3) received more than 7 different answers.

4 Evaluation of the paraphrase corpus

The final corpus was assessed by three philologists in order to find real paraphrases among the candidates.

From every group of related statements in the corpus, all the possible pairs of statements among those of the group were considered for evaluation. Thus, from a group of m related statements, $C_{m,2} = \binom{m}{2}$ pairs must be evaluated. For the present case, 393 pairs were produced for evaluation.

The assessors were asked to consider the context of the statements and to admit some redundancies between the affirmative form of the question and its answers. For example, for the affirmative form of the question *"¿Qué es el Atlantis?"* (What is Atlantis?), that is *"Atlantis"*, four different answers are associated:

1. *"transbordador estadounidense"* (american shuttle)
2. *"foro marítimo"* (marine forum)
3. *"transbordador espacial atlantis"* (space shuttle)
4. *"transbordador espacial estadounidense"* (american space shuttle)

As it can be observed, the answer *"foro marítimo"* does not pertain to the same context than the other answers, but *"Atlantis"* and *"foro marítimo"* were considered a paraphrase, such as *"Atlantis"* and *"transbordador espacial estadounidense"*. But *"foro marítimo"* and *"transbordador espacial estadounidense"* were not, obviously, considered a paraphrase. About redundancies, it can be observed that *"transbordador espacial atlantis"* contains *"Atlantis"*, but both statements express the same idea, i.e., they are a semantic paraphrase. In addition, this example illustrates the affirmation given by (Shinyama et al., 2002) that expressions considered as paraphrases are different from domain to domain.

The evaluators labeled every single pair with a boolean value: YES if it was considered that a paraphrase was given between both statements, and NO on the contrary. The assessments of the three experts were used as a votation. Then, for every possible pair of statements, it was finally decided that it was a paraphrase if at least two of the labels given by the assessors to the pair were YES. Following this criterion, from the 393 candidate pairs of statements, 291 were considered paraphrases, i.e., 74%. The agreement inter–annotator was of 76%. The three experts labeled simoultaneously with YES 204 pairs, and labeled simoultaneously with NO 48 pairs. Then, a total agreement was given for 252 pairs, i.e., 86.6% of the ones that were considered paraphrases.

4.1 Problems detected and suggestions for improvement

The biggest disagreements between annotators were given in "difficult" pairs such as, for example: *"países que forman la OTAN actualmente"* (countries that conform the NATO at the moment) and *"dieciséis"* (sixteen); this is because, for some people, a number can not substitute a set of countries but, for some other people, in a determined context it can be said, indifferently, for example: *"... the countries that conform the NATO at the moment held a meeting in Paris last week..."* or *"... the sixteen held a meeting in Paris last week..."*.

This situation suggested the analysis of the pairs involved in disagreements. From it, several phenomena were detected. The most frequent ones are shown in the following list:

- Some errors are introduced by the annotators, because they do not consider accurately the context in which the statements are. As an example, one of the annotators did not consider the pair *"organización que dirige yaser arafat"* (organization leaded by yasser arafat) and *"autoridad nacional palestina"* (palestinian national authority) a paraphrase because nowadays Yasser Arafat does not lead the Palestinian National Authority.
- When one of the statements of the pair comes from a factoid–type question of the QA exercise, and its answers are restricted to a date (see (Magnini et al., 2006) for more information about this kind of questions and answer restrictions), then "difficult" pairs as the following appear: *"murió stalin"* (stalin died) and *"5 de marzo de 1953"* (5th March 1953). Some annotators consider that there is a paraphrase but it is because they infer some words that are missing in the affirmative form of the question in order to complete the overall context of the pair. Thus, for this pair some annotators actually understand *"fecha en la que murió stalin"* (date in which stalin died) instead of *"murió stalin"*. This example shows that some disagreements can be induced by the transformation into affirmative form.
- Some annotators are more strict than others when considering the grammatical accuracy of the statements. QA systems sometimes introduce little grammatical errors in their responses, and this affects the consideration about the existence of paraphrase. This is more frequent in answers given to date–type or location–type questions, because of the format given to them by the QA systems. The following two examples illustrate the case: first, in the pair *"3 de abril de 1930"* (3^{rd} april 1930) and *"3 abril 1930"* (3 april 1930), the first statement is correct but in the second the prepositon *"de"* is missing; despite the fact that it can be perfectly understood, some annotators think that it has no sense; second, in the pair *"lillehammer (noruega)"* (lillehammer (norway)) and *"lillehammer noruega"* (lillehammer norway), the lacking parentheses in the latter statement made some annotators consider that it could be interpreted as a compound name instead of a pair of names (the city and its country).
- Another source of disagreement is the fact that there is not a bidirectional entailment between the two statements of the pair. The pair *"lepra"* (leprosy) and *"enfermedad infecciosa"* (infectious disease) serves as an example. Leprosy is a infectious disease, but not every infectious disease is leprosy. Despite of this fact, some annotators considered that there is a paraphrase, because under determined contexts both statements can be used indifferently.
- Sometimes, errors acquired from the QA assessment process cause different opinions among the annotators. For example, the pair *"deep blue"* and *"ordenador de ajedrez"* (chess computer) is in the corpus because the assessors of the QA exercise considered *"ordenador de ajedrez"* (chess computer) as an adequate answer for the question *"¿qué es deep blue?"* (what is deep blue?). Despite the fact that the annotators were asked to consider all the statements as validated, those of them who knew that, in fact, Deep Blue is not a computer devoted to play chess, did not label the pair as paraphrase.

These problems suggest that the assess-

ment process should be improved. Thus, not only a simple labelling action but a more complex process should be accomplished. Two alternative propositions for a better assessment process are outlined here:

1. In a first round, the assessors not only label the pairs but write an explanation for every decission. In a second round, independent assessors take a definitive decision having into account both the votation among the labels given in the previous round and the considerations written.
2. In a first round, the assessors only label the pairs and, in a second round, they discuss the controversial cases, and everyone can reconsider its opinion to relabel the pair; if an agreement is not reached, the pair and the opinions are submitted to independent assessors.

In addition, the assessment process should be supervised in order to homogenize criteria about what kind of little errors should be considered by the assessors; for example, the lack of parentheses of prepositions.

Of course, some errors can not be avoided when applying a fully automated process. For example, pairs without sense such as *"deep blue"* and *"ordenador de ajedrez"* (chess computer), that depend on the QA assessment process, can not be identified with shallow techniques.

4.2 What do humans understand under paraphrase?

Several methods for recognizing paraphrases or obtaining them from corpora have been proposed until now, but a doubt arises: what is exactly what these methods are recognizing or obtaining? The definition for paraphrase is very fuzzy and context–dependant, as seen here; even more, almost every author gives a definition of his own; for example, the one given by (Fabre and Jacquemin, 2000):

> *Two sequences are said to be a paraphrase of each other if the user of an information system considers that they bring identical or similar information content.*

Or the one by (Wan et al., 2006):

> *[...] paraphrase pairs as bi–directional entailment,*

where a definition for entailment can be found in (Dagan et al., 2006):

> *Entailment: whether the meaning of one text can be inferred (entailed) from the other.*

But these and the other definitions that can be found for paraphrase can be included in the simple concept given by (Shinyama et al., 2002):

> *Expressing one thing in other words.*

This last enunciation is very useful because it is capable to deal with the variety of human opinions. But it is not restrictive at all. The difficulty when working with paraphrases lies on its own definition. This is because of the relatively poor agreement when different persons have to say if a pair of expressions can be considered paraphrases. Thus, paraphrase corpora could be built or paraphrase recognition systems could be developed, but every single system using such resources should be capable of discriminating the usefulness of the supplied sets of paraphrases.

5 Conclusions and future work

The annotated corpora from the assessment processes of campaigns like the CLEF, the TREC or the NTCIR, grow year by year. This human work generates a great amount of validated data that could be successfully reused. This paper describes a very simple and little costly way to obtain paraphrases is described, but it is ot the only nor the more complex issue that can be accomplished. Thus, corpora –aimed at different applications– could be increased every year using the newest results of this kind of campaigns. In addition, the rules proposed here for transforming questions into their affirmative form can be used for automatically building the corpora needed in future AVEs.

Despite the fact that the concept of paraphrase is human–dependant and, therefore, it is not easy to obtain a high agreement inter–annotator, it has been showed that a high amount of paraphrases can be obtained by means of shallow techniques. Anyway, the assessment process applied to the paraphrase candidates corpus can be improved; several ideas for this have been outlined in this paper. As a result of this improvement, the agreement inter–annotator should increase and the percentage of identified paraphrases should decrease, but hopefully not to the point in which the proposed method should be considered useless. In the near future new models for this assessment process should be

evaluated, in order to determine the most appropriate one. Appart from the accuracy of the assessment process, the results obtained at the present time suggest that it will be interesting to test if paraphrase corpora, as the one presented in this paper, are really useful for different applications; and if it is worthwhile to implement more complex techniques or the little errors produced do not interfere with the performance of these applications. This will determine if such corpora should be obtained every year after evaluation campaings as the one accomplished at CLEF.

References

R. Barzilay and K.R. McKeown. 2001. *Extracting Paraphrases from a Parallel Corpus.* Proceedings of the ACL/EACL.

W. Bosma and C. Callison–Burgh. 2006. *Paraphrase Substitution for Recognizing Textual Entailment.* Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain.

Ido Dagan, Oren Glickman and Bernardo Magnini. 2006. *The PASCAL Recognising Textual Entailment Challenge.* MLCW 2005. LNAI. Springer. 3944, Heidelberg, Germany.

Cécile Fabre and Christian Jacquemin. 2000. *Boosting Variant Recognition with Light Semantics.* Proceedings of the 18th conference on Computational linguistics - Volume 1, Saarbrucken, Germany.

Junichi Fukumoto, Tsuneaki Kato and Fumito Masui. 2004. *Question Answering Challenge for Five Ranked Answers and List Answers – Overview of NTCIR4 QAC2 Subtask 1 and 2 –.* Working notes of the Fourth NTCIR Workshop Meeting, National Institute of Informatics, 2004, Tokyo, Japan.

Tsuneaki Kato, Junichi Fukumoto and Fumito Masui. 2004. *Question Answering Challenge for Information Access Dialogue – Overview of NTCIR4 QAC2 Subtask 3–.* Working notes of the Fourth NTCIR Workshop Meeting, National Institute of Informatics, 2004, Tokyo, Japan.

D. Lin and P. Pantel. 2001. *Discovery of Inference Rules for Question Answering.* Natural Language Engineering, 7(4):343–360.

Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Petya Osenova, Anselmo Peñas, Valentin Jijkoun, Bogdan Sacaleanu, Paulo Rocha and Richard Sutcliffe. 2006. *Overview of the CLEF 2006 Multilingual Question Answering Track.* Working Notes of the CLEF 2006 Workshop, 20–22 September, Alicante, Spain.

B. Pang, K. Knight and D. Marcu. 2003. *Syntax–based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences.* NAACL–HLT.

A. Peñas, Á. Rodrigo, V. Sama and F. Verdejo. 2006. *Overview of the Answer Validation Exercise 2006.* Working Notes of the CLEF 2006 Workshop, 20–22 September, Alicante, Spain.

D. Ravichandran and E. Hovy. 2002. *Learning Surface Text Patterns for a Question Answering System.* Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).

Y. Shinyama, S. Sekine, K. Sudo and R. Grishman. 2002. *Automatic Paraphrase Acquisition from News Articles.* Proceedings of HLT, pages 40–46.

E.M. Voorhees and H.T. Dang. 2005. *Overview of the TREC 2005 Question Answering Track.* NIST Special Publication 500–266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), Gaithersburg, MD, USA.

Stephen Wan, Mark Dras, Robert Dale, and Cecile Paris. 2006. *Using Dependency–Based Features to Take the "Para–farce" out of Paraphrase.* Proceedings of the Australasian Language Technology Workshop 2006, Sydney, Australia.

Yitao Zhang and Jon Patrick. 2005. *Paraphrase Identification by Text Canonicalization.* Proceedings of the Australasian Language Technology Workshop 2005, Sydney, Australia.