Procesamiento del Lenguaje Natural, Revista nº45, septiembre 2010, pp 11-19 recibido 30-04-10 revisado 17-05-10 aceptado 20-05-10

WRPA: A System for Relational Paraphrase Acquisition from Wikipedia*

WRPA: Un sistema para la adquisición de paráfrasis de relaciones de la Wikipedia

Marta Vila
CLiC-UB
Gran Via 585, Barcelona
marta.vila@ub.edu

Horacio Rodríguez
TALP-UPC
Jordi Girona 1-3, Barcelona
horacio@lsi.upc.es

M. Antònia Martí
CLiC-UB
Gran Via 585, Barcelona
amarti@ub.edu

Resumen: En este artículo se presenta WRPA, un sistema para la Adquisición de Paráfrasis de Relaciones de la Wikipedia. Aprovechando la estructura de la Wikipedia, WRPA extrae patrones de paráfrasis que expresan una determinada relación entre dos entidades. La novedad de este sistema reside en que se explota dicha enciclopedia más allá de las fichas (o infoboxes), aprovechando información itemizada que contienen algunas de sus páginas. WRPA es independiente de la lengua, asumiendo la existencia, para la lengua en cuestión, de Wikipedia y de herramientas para el tratamiento superficial del lenguaje, así como independiente de la relación tratada.
Palabras clave: Paráfrasis, Extracción de Información, Extracción de Relaciones, Wikipedia.

Abstract: In this paper we present WRPA, a system for Relational Paraphrase Acquisition from Wikipedia. WRPA extracts paraphrasing patterns that hold a particular relation between two entities taking advantage of Wikipedia structure. What is new in this system is that Wikipedia's exploitation goes beyond infoboxes, reaching itemized information embedded in Wikipedia pages. WRPA is language independent, assuming that there exists Wikipedia and shallow linguistic tools for that particular language, and also independent of the relation addressed.
Keywords: Paraphrasing, Information Extraction, Relation Extraction, Wikipedia.

1 Introduction

Paraphrasing stands for (approximate) sameness or equivalence of meaning between different wordings. This definition puts into words a vague and complex phenomenon with a broad range of manifestations that can involve lexical, syntactic, semantic and pragmatic knowledge. NLP components dealing with paraphrasing appear to have great potential for the improvement of understanding and generation systems such as question-answering, summarization or machine translation. As a result, it has been the focus of a large amount of work in the last couple of decades.

In this paper we present WRPA, a system for Relational Paraphrase Acquisition from Wikipedia. Due to the vagueness and complexity of the paraphrasing phenomenon, we restrict ourselves to relational paraphrases, i.e., those expressing a relation between two entities, because they constitute a well delimited but in turn comprehensive set.

Our approach to paraphrasing has a close relationship with Information Extraction systems, as they are frequently used for extracting semantic relations. However, while IE systems are geared towards obtaining the semantic relation held by pairs of entities—named the source and the target—in a corpus (Figure 1), paraphrasing focusses on the wording used to express those relations (patterns and instances in Figure 1). A lot of techniques can be used in IE, e.g., machine learning and rule- or pattern-based techniques. WRPA is only related to the latter.

Our approach is based on Harris (1954)'s Distributional Hypothesis which states that

* This work is supported by the FPU Grant AP2008-02185 from the Spanish Ministry of Education, and the Text-Knowledge 2.0 (TIN2009-13391-C04-04) and KNOW2 (TIN2009-14715-C04-04) projects.

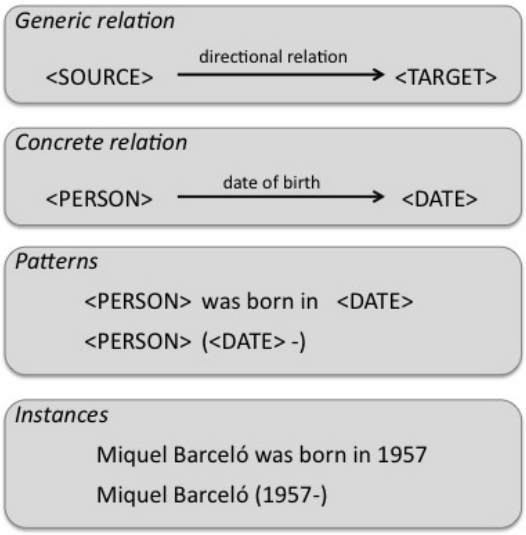

Figure 1: Relation Extraction

words occurring in the same contexts tend to have similar meanings. Our 'contexts' are the source and target entities, and the texts around them are our candidate paraphrases. WRPA automatically obtains source and target entities for a concrete relation taking advantage of Wikipedia structure and uses them as anchor points for the acquisition of candidate paraphrasing patterns. What is new in our system is that it takes advantage of Wikipedia structure beyond infoboxes, using itemized information embedded in Wikipedia pages. This allows for the extraction of a large number of highly reliable source and target entities.

In this paper, we focus on the authorship relation understood in a broad sense, i.e. including several sub-relations. In our approach, the category authors (source) corresponds to different professional or artistic activities and includes painters, sculptors, architects, writers, composers, singer-songwriters, directors, philosophers, inventors and scientists. The works resulting from their activity (target) include paintings, sculptures, constructions, books, articles, musical compositions, albums, films, theories, inventions, etc. This variety guarantees the appropriateness of this relation for research in paraphrasing, since it provides a rich casuistic of paraphrases. Moreover, a system dealing with such a complex relation can be easily transported to other (simpler) relations. In this paper we also deal with date of birth, date of death and place of birth relations, the source being a person and the target being the birth or death information. Henceforth, these relations will be referred to as 'person' relations, as opposed to the 'authorship' relation.

Finally, this method is language independent, assuming that there exist both Wikipedia and shallow linguistic tools for that particular language. In order to demonstrate the portability of our system, the authorship relation is handled using the Spanish Wikipedia; person relations, in turn, are addressed using the English Wikipedia.

2 Related Work

Different approaches have been developed for paraphrase extraction from corpora. From the perspective of the preprocess needed, some systems work with parsed data. DIRT (Lin and Pantel, 2001) extracts paraphrases based on an extended version of Harris's Distributional Hypothesis. This hypothesis states that if two paths in a dependency tree tend to occur in similar contexts, the meaning of the paths tends to be similar. TEASE (Szpektor et al., 2004), in turn, departing from lexicon entries, extracts anchor sets for those entries and, departing from sentences containing these anchor sets, extracts paraphrases for the initial entries.[1] As these systems require syntactic parsing, they present a limitation regarding the number of languages susceptible to their application.

Other systems rely on shallow linguistic processing. Brin (1998), Agichtein and Gravano (2000) and Ravichandran and Hovy (2002) apply bootstrapping approaches. Their basic strategy is, starting with some seed entity pairs that hold a particular relation, to extract expressions containing those pairs and use them to find new pairs iteratively. Bhagat and Ravichandran (2008), in turn, depart from a few seed patterns for a target relation to extract new patterns expressing this relation. A drawback of these approaches is the fact that they need to be provided with some highly reliable initial seeds. Moreover, methods that apply bootstrapping normally need a careful control in order to avoid degradation in the results obtained. Some of these authors have already dealt with the authorship relation, but from

[1] It has to be said that these systems, broadening the paraphrasing scope, are geared towards finding entailment templates.

a narrower perspective, e.g. author-book or inventor-invention relations.

Sekine (2005), without requiring any initial seed, carries out NE pair matching for paraphrase extraction in newswire corpora. Barzilay and Lee (2003) perform argument matching in the same type of corpus, but they extract paraphrases taking advantage of the comparison between articles reporting the same event. A drawback of these systems is the questionable quality and appropriateness of the NEs or the arguments used as anchor points to extract paraphrases.

Monolingual parallel corpora have also been exploited. Barzilay and McKeown (2001) use different translations of the same literary text and apply a co-training algorithm that takes advantage of words and their context to extract paraphrases. Zhao et al. (2008; 2009) use bilingual parallel corpora to extract paraphrases in English using the sentences in another language as pivots. These approaches present some limitations regarding corpora availability, as monolingual and bilingual parallel and comparable corpora are limited in number.

Our work is also related to relation extraction in IE (Turmo, Ageno, and Català, 2006). Most approaches use supervised machine learning techniques. The key point of such approaches is the need for human supervision, which is costly and time consuming. Many techniques have been applied to reduce this cost from bootstrapping to active or semi-supervised learning. A recent approach consists of adapting highly accurate structured resources for providing learning material. Wikipedia is recognized as an exceptional resource in the NLP community (Medelyan et al., 2009) and it has been used extensively for extracting lexical and conceptual information. Using infoboxes for obtaining cheap supervised examples for learning information extractors was first proposed by Wu and Weld (2007) and Wu, Hoffmann, and Weld (2008), and later by Gokalp et al. (2009). A problem that arises with these approaches is the limited coverage and the variablity of the infobox content.

3 The Corpus

Our corpus consists of the Spanish and English versions of Wikipedia (henceforth WP, SWP and EWP).[2] We downloaded the SWP and EWP versions of February 2009 into a MySQL database. SWP comprises 484,550 pages (3.1 GB of textual content) and EWP, 1,660,067 pages (12.8 GB).

Extracting information from WP can be done easily using a Web crawler and a simple HTML parser. The structured format of WP pages allows for this simple procedure. There are, however, a lot of APIs providing easy access to WP online or to the database organized data obtained from WP dumps. For accessing the database we used Iryna Gurevych' JWPL[3] software.

WP information unit is the 'Article' (or 'Page'). The set of articles and their links in WP form a directed graph. Moreover, every article can be assigned to one or more WP categories through 'Category links'. At the same time, a category is linked to one or more categories (super- and sub-categories), structuring themselves as classes that are also organized as a graph. There are several types of special pages in WP: 'Redirection pages', i.e., short pages which often provide equivalent names for an entity, and 'Disambiguation pages', i.e., pages with little content that link to multiple similarly named articles.

Figure 2: SWP infobox

WP articles are semistructured, i.e., they contain structured information and free text. Structured information consists of tagged parts following templates (infoboxes) or itemized sections. An infobox provides a summary of the information within articles and it consists of attribute-value pairs. Attributes can be univalued or multivalued (*fecha de nacimiento*, 'date of birth', and *obras desta-*

[2] http://www.wikipedia.org

[3] http://www.ukp.tu-darmstadt.de/software/jwpl/

cadas, 'outstanding works', respectively, in Figure 2). Itemized sections, in the case of the authorship relation in our experiments, contain an itemized list of works by the author heading the article. There also exist WP pages only containing itemized works by a concrete author.

The structure of WP provides us with a good starting point for automatically extracting paraphrasing patterns. We exploit infobox attributes and their values, and itemized sections and pages in order to extract our target entities. Our source entities are extracted from article titles and redirection pages. WP category structure is also exploited in our system.

4 Methodology

Our hypotheses are i) following Harris, the meaning of the text around the source and target entities will be similar throughout their different ocurrences; and ii) this meaning will hold the source-target relation in some way. Thus, WRPA uses source and target entities as anchor points for our candidate paraphrase extraction.

Candidate paraphrasing patterns extracted by our system present the following structure:

{text} [X] {text} Y {text} [Z] {text}

with X, Y and Z anchor points occurring in any order and only Y being mandatory. X stands for the source (author and person in our experiments) and Y stands for the target (work, and birth or death information). In the case of the authorship relation, Z includes the work creation year, which, when appearing, makes the pattern stronger and more reliable. For the person experiments, only X and Y are considered.

Our system can be divided into three steps, reflected in Figure 3. First, the corpus is set up (Section 4.1). Second, X, Y and Z anchor points are extracted (Section 4.2). Third, the candidate paraphrasing patterns are obtained (Section 4.3). In our experiments, grey sections in Figure 3 only apply to the authorship relation.

4.1 Corpus set up

We collect the WP articles that correspond to the source. For the authorship relation experiment, as a category including all authors does not exist in the SWP, we had to

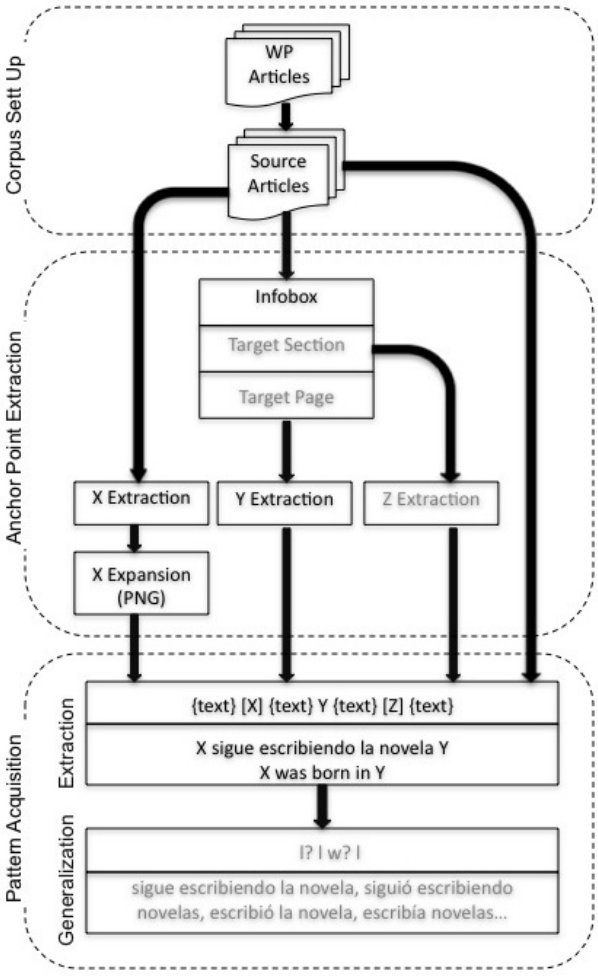

Figure 3: WRPA scheme

follow a multiple-step approach. First, we manually browsed the infoboxes under the infobox category 'people'[4] and then we selected all infoboxes containing a 'work' attribute. We carried this out manually because the concept of 'work' is expressed in many different ways in the SWP (e.g. *obras destacadas*, 'outstanding works' or *sencillo exitoso*, 'hit single'). This resulted in a set of 22 infobox/attribute pairs[5] (e.g. the pair *Ficha de Artista/obrasdestacadas*—'Artist Infobox/outstanding works'—in Figure 2). Then, we automatically extracted all the categories that contained articles including these infoboxes. This resulted in a set of 4,054 categories. Finally, we recovered all the articles—including or not an infobox—

[4] `http://es.wikipedia.org/wiki/Categoría:Wikipedia:Fichas_de_personas`

[5] It has to be taken into account that a single infobox can have several attributes referring to 'work' and that a single attribute can apply to several infoboxes.

contained in these categories. This collection of 47,466 SWP pages constitute our working subset.

For the person experiments, we followed the same methodology. However, as we took advantage of the mappings from generic attributes to specific ones provided for the KBP contest[6] (e.g. the *date of birth* generic attribute is mapped to specific attributes like *datebirth* or *birth date*), we were able to perform this step without human intervention. We started with 6 generic attributes covering date of birth and death, and place of birth relations, which were mapped to 25 specific attributes. This resulted in a set of 302 infobox/attribute pairs, 20,741 categories, and 134,902 pages containing date of birth, 35,755 pages containing date of death and 91,739 pages containing place of birth.

The working subsets were then preprocessed. Each page was cleaned (non valid encoded material, and HTML, XML and Wikimedia tags were removed), normalized and segmented into sentences.

4.2 Anchor Point Extraction

The extraction of X and Y anchor points is done independently and through different processes. Z (only for authorship) is obtained alongside Y. We extract the X anchor points from the titles of the source pages as well as the redirection pages. We extract the Y and Z anchor points from infoboxes, and target sections and pages.

For the person experiment, information extracted from the target attribute values in infoboxes was enough for Y anchor point extraction, because i) it corresponds to univalued attributes, ii) the EWP is very extensive, iii) there is a relatively high number of infoboxes for person pages (from the 419,058 person pages, 142,452 contained infoboxes, i.e., 34%) and iv) birth and death information generally appears in person infoboxes.

In the case of the authorship experiment, the initial Y anchor points were also gathered from infoboxes. Z anchor points were only collected when they appeared next to Y in infoboxes. However, the coverage of Y and Z was usually low, because i) the SWP is smaller than the EWP, ii) most author pages lack an infobox, iii) most infoboxes lack the work attribute and iv) infoboxes contain only the most important works (on average two per infobox). Moreover, multivalued attributes like 'works' consist of a string with possible noise and, thus, are less reliable than univalued ones. In order to improve the recall, we took advantage of both the work section and the links to specific pages containing works. Work sections are less reliable than infoboxes but allow for the extraction of many examples. Problems presented by the work section can be summarized as i) the difficulty involved in obtaining the section limits (heading, ending) and ii) the difficulty involved in extracting the values because of the presence of complementary material such as suffixes or prefixes. Work pages are highly reliable as the above problems do not appear.

By way of illustration, in the case of Paul Auster, the existing infobox lacks the 'work' attribute. In contrast, 33 works (of which 25 were correct) were extracted from the work section and 18 more (all correct) were added from the work page.

Finally, information encoded in the infobox attribute values presents some variablity as shown in Figure 4 for work and date of birth attributes. In order to deal with this variablity, we attach a small manually written[7] CFG (30 productions on average) to each attribute type. The infobox attribute value is then tokenized and parsed to obtain Y and eventually Z anchor points.

```
| obrasdestacadas : ''[[Las señoritas de Avignon]]'' (1907)
<br />''[[Guernica (cuadro)|Guernica]]'' (1937)

| obrasdestacadas : ''Mural Capitolio'' ''Interludio''

|birth_date : [[August 15]], [[1974]]

|birthdate : 1960|8|24
```

Figure 4: Infobox attributes and their values

Table 1 contains some anchor point examples extracted by our system.[8]

4.3 Paraphrasing Pattern Acquisition

Once the set of anchor points is defined, we can go on to the paraphrasing pattern acquistion. It consists of two steps: the extraction of candidate paraphrasing patterns

[6] http://nlp.cs.qc.cuny.edu/kbp/2010/

[7] An automatic learning of the grammar, with or without manual revision, could be considered instead, and will probably be used in the future for scaling up our method.

[8] We do not include 'date of death' examples, as they are parallel to 'date of birth' ones.

X	Y(Z)
Authorship	
Canaletto	La Riva degli Schiavoni (1730-31)
Edgar Allan Poe	[[Eureka]]
Luis Eduardo Aute	Templo de carne (1986)
Date of birth	
David Kaye	[[October 14]], [[1964]]
Sara Rue	1979|01|26
Joan of Arc	c. [[1412]]
Place of birth	
Giovanni Branca	[[San Angelo]] in [[Lizzola]], [[Pesaro]]
Grigore Preoteasa	[[Bucharest]], [[Romania]]
Tomas Plekanec	[[Kladno]], [[Czech Republic |CZE]]

Table 1: Anchor point examples

in the body of WP articles using X, Y and Z as anchor points (Section 4.3.1) and pattern generalization (Section 4.3.2). In our experiments, the latter only applies to the authorship relation.

4.3.1 Candidate Paraphrasing Pattern Extraction

Candidate paraphrasing pattern extraction consists of looking for the obtained X, Y and Z anchor points in the body of the articles and gathering them together with the text around them occurring between period markers.

In order to improve the recall we look for all the variant forms that express X and Y. For instance, for *Paul Auster*, variants include cases like *Auster* and *P. Auster*. For *April 3 1947*, variants include *1947* and *April, 1947*, among others. In order to expand X, we apply a proper name grammar that formalizes the different ways in which the proper name of a person can be expressed (Arévalo, Civit, and Martí, 2004) to each WP variant.[9] In the case of person experiments, the expansion of Y is performed by generalizing and filtering the parsed trees obtained when parsing the infobox attribute value, and using the same grammar for generation.

Several types of pattern candidates were extracted depending on the presence and order of X, Y and Z. In the following, we will concentrate on <X text Y> patterns. When applying these patterns to new corpora, we establish the Y right limit in the longest snippets of text satisfying the grammar.

[9] Although initially developed for Spanish the grammar has been adapted to English.

Table 2 shows examples of <X text Y> patterns extracted by our system. The patterns contain some variables. By way of illustration, the variable *YEAR* in 3 stands for a Z that is not linked to the Y in the pattern. This can be because it corresponds to another Y in our anchor set (case of 3), or because the system has not detected the relation between them. The variable *PERSON* in 4 and 9 stands for a name variant of X, and, as can be seen in 5, it gives rise to another pattern.

	Authorship
1	X sigue escribiendo la novela Y *X continues writing the novel Y*
2	X comenzó a grabar su álbum debut, "Y" *X started to record his debut album, "Y"*
3	X dirigió "Educando a Rita" (YEAR) y "Y" *X directed "Educando a Rita" (YEAR) and "Y"*
	Date of birth
4	X known as PERSON was born in Y
5	X was born in Y
6	X was born in PLACE on Y
	Date of death
7	X DATE-Y
8	X DATE to Y
9	X PERSON DATE - Y
	Place of birth
10	X born DATE in Y
11	X DATE in Y
12	X DATE Y

Table 2: Pattern examples

4.3.2 Candidate Paraphrasing Pattern Generalization

In the case of person experiments no further processing is needed. The obtained patterns are straightforwardly transformed into regular expressions and their precision and recall are computed. Precision is computed by fixing X and then applying the pattern to the original page in order to obtain all possible Ys. If the resulting Y is a variant or the original Y extracted from the infobox the agreement is positive. For estimating recall, we make the conservative assumption that all the pages contain a date of birth, a date of death and a place of birth. Obviously, this is not always the case (e.g. assuming the existence of 'date of death' is not pertinent for living people).

Regarding authorship, candidate pattern generalisation implies going from a representation of the pattern as a sequence of words to its representation as a sequence of tokens that

take the form of words, lemmas or PoS tags. Additionally, a token can be mandatory, skippable or omitted. The final form of the generalized pattern will be transformed into a regular expression formula. Generalization aims to extract common patterns from our initial ones. These common patterns will explain a larger number of texts.

In the generalization process, we first PoS tag candidate patterns using the Freeling toolbox.[10] Then, we represent each pattern as a sequence of <word, lemma, PoS> tuples (Table 3).[11]

word	sigue	escribiendo	la	novela
lemma	seguir	escribir	el	novela
PoS	vm	vm	da	nc

Table 3: Token-based pattern examples

The process of generalization of a pattern is performed using an A* approach until N[12] matches with other patterns are reached. Each state is represented as a sequence of tuples consisting of a token (word, lemma or PoS) and a condition (Table 4).

w	The token has to match the word
w?	The token has to match the word or be skipped
l	The token has to match the lemma
l?	The token has to match the lemma or be skipped
p	The token has to match the PoS
p?	The token has to match the PoS or be skipped
-	The token must be skipped

Table 4: Token matching conditions

In the initial state, all the tokens fulfil the **w** condition. The operators allow for moving from **w** to **w?**, from **w?** to **l** and so on. The cost of moving from one state to another depends on the operation (e.g. moving from **w** to **w?** has a cost of 1, from **w?** to **l** a cost of 0.7 and so on).

As the algorithm suffers from the 'plateau' problem (for many states, especially in the first state of the searching, the heuristic function is the same), we have implemented several macrooperators that allow for the exploration of distant areas of the search space. In order to reduce the search space, we have performed a clustering of the candidates so that

[10] `http://www.lsi.upc.edu/~nlp/freeling/`

[11] vm (verb, main), da (definite, article), nc (noun, common).

[12] Being set to 10 in our experiments.

Initial state	<sigue:w> <escribiendo:w> <la:w> <novela:w>
Matching pattern	sigue escribiendo la novela ('continues writing the novel')
Generalization	<sigue:l?> <escribiendo:l> <la:w?> <novela:l>
Matching patterns	1) sigue escribiendo la novela 2) siguió escribiendo novelas ('continued writing novels') 3) escribió la novela ('wrote the novel') 4) escribía novelas ('wrote novels')

Table 5: Pattern generalization

pattern generalization is carried out within each cluster. The distance measure used for clustering is a simple Levenshtein distance between the string of lemmas in each pattern.

In the authorship experiment, precision is computed in the same way as in person experiments, i.e., comparing the works obtained by applying our patterns to our departure anchor set. However, in this case, there are no clear alternatives for automatically measuring the recall: it is not possible to know how many works occur in the WP pages, so we do not provide this measure.

5 Experiments and Results

Two experiments were carried out in order to test the performance of WRPA: the 'authorship' experiment, using the SWP, and the 'person' experiment, using the EWP.

In the authorship experiment, we processed the set of 47,466 SWP pages obtained in the corpus set up (Section 4.1). Only 8,343 had an infobox and, among them, only 305 had a 'work' attribute. In contrast, 17,715 pages had a work section and 593 had a link to a work page. This resulted in obtaining 575 author-work pairs from the infoboxes and 233,376 pairs from the work sections and pages. From these pairs, we obtained 32,288 candidate paraphrasing patterns containing at least a Y.[13]

Regarding the person experiment, 419,058 EWP pages were processed. From them, 142,452 contained an infobox, 154,845 containing the date of birth attribute, 35,755 the

[13] The subject (normally the X in our patterns) elision is extremely frequent in Spanish.

			Precision	Recall	F1
Person	**Date of birth**	X born Y	0.95	0.57	0.71
		X Y	0.80	0.12	0.21
		Top 8 patterns	0.92	0.75	0.83
		Baseline	0.95	0.57	0.71
	Date of death	X DATE-Y	0.95	0.21	0.34
		X PERSON DATE-Y	0.96	0.10	0.18
		Top 3 patterns	0.95	0.42	0.58
		Baseline	0.95	0.21	0.34
	Place of birth	X born DATE in Y	0.98	0.13	0.23
		X DATE in Y	0.94	0.10	0.18
		Top 3 patterns	0.92	0.26	0.41
		Baseline	0.98	0.13	0.23
Authorship	<pintó:w><su:w?><cuadro:w?>		0.46	–	–
	<pintó:w><su:w?><primera:l?><obra:w?>		0.49	–	–

Table 6: Results

date of death attribute and 91,739 the place of birth attribute. This resulted in 161,398 candidate patterns for date of birth, 36,895 for date of death and 93,021 for place of birth.

In the person experiment, the baseline is established by the most frequent pattern extracted by our system. Using the same criteria for the authorship relation makes no sense because of the extremely high number of patterns presenting a very low recall. Furthermore, attempts to artificially build a baseline resulted in very low precision. We leave for the future further research in order to determine whether an authorship baseline is feasible or not.

The results of our system are collected in Table 6. The first couple of patterns in each person relation are the most frequent. In the case of date of birth, selecting the top 8 patterns resulted in a high recall, clearly outperforming the baseline. In the other cases, the recall is low due to the conservative criteria used for defining it. The figures, however, clearly outperform the baseline.

Precision for the patterns in the person experiment is extremely high, due to the low level of difficulty of the task as the number of patterns is very limited. In the case of the authorship experiment, there are many patterns with similar accuracy. We present a couple of them having a good coverage. The precision in these cases is lower due to the difficulty of the task, but high enough for our purposes, i.e., submitting to human judgement the paraphrasehood of the candidates (future work in Section 7).

6 Conclusions

WRPA is a new system that takes advantage of Wikipedia structure in order to obtain relational paraphrase patterns. Our system overcomes some drawbacks present in previous work. First, WRPA only relies on WP and shallow linguistic processing: lemmatizing and PoS tagging is needed for the generalization step. Second, finding good contexts is essencial within paraphrasing systems based on the Distributional Hypothesis, as it is on the quality of these contexts that the quality of the obtained paraphrases depends. We guarantee the quality of these contexts, i.e., our source and target entitites, as they are directly extracted from structured and semantically labelled data. Thus, external NE tagging is not needed. Third, going beyond infoboxes and using itemized information embedded in WP pages, allows for the acquisition of a large number of initial contexts, which makes unnecessary the application of bootstrapping techinques.

Our experiments prove that our system is language independent—WP has articles in more than 250 languages—and also independent of the relation to be extracted.

7 Future Work

In the generalization process, we have established the A* matches as well as the transition costs arbitrarily. In future work, we aim to conduct a study of the cases in order to establish new and justified parameters. In the case of the transition costs, we plan to take advantage of España-Bonet et al. (2009),[14]

[14] `http://www.lsi.upc.edu/~textmess/`

a collaborative web interface for the compilation of paraphrases. This interface will also be used in the future to manually validate (i.e., judge their paraphrasehood) the patterns obtained by WRPA.

Moreover, further work will be carried out in order to see whether an authorship baseline is feasible or not. Also, a more serious evaluation of the recall as well as an evaluation on non-WP texts will be undertaken.

Finally, this system will be tested in the KBP contest (note 6) where it will be applied to 26 different relations in the EWP.

References

Agichtein, Eugene and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collocations. In *Proceeding of ACM DL 2000*, pages 85–94.

Arévalo, Montserrat, Montserrat Civit, and M. Antònia Martí. 2004. Mice: A module for named entity recognition and classification. *International Journal of Corpus Linguistics*, 9(1):53–68.

Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*, pages 16–23.

Barzilay, Regina and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the ACL 2001*, pages 50–57.

Bhagat, Rahul and Deepak Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of the ACL 2008*, pages 674–682.

Brin, Sergey. 1998. Extracting patterns and relations from world wide web. In *Proceeding of WebDB 1998*, pages 172–183.

España-Bonet, Cristina, Marta Vila, M. Antònia Martí, and Horacio Rodríguez. 2009. Coco, a web interface for corpora compilation. *Procesamiento del Lenguaje Natural*, (43):367–368.

Gokalp, Sedat, Syed Toufeeq Ahmed, Suvitha Vijayarajan, and Hasan Davulcu. 2009. Wikisld: Mapping wikipedia infobox information onto the article text. In *Proceeding of WikiAI09 Workshop of IJCAI 2009*.

Harris, Zellig. 1954. Distributional structure. *Word*, 10(23):146–162.

Lin, Dekang and Patrick Pantel. 2001. Dirt-discovery of inference rules from text. In *Proceedings of ACM SIGKDD 2001*, pages 323–328.

Medelyan, Olena, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from wikipedia. *International Journal of Human-Computer Interactions*.

Ravichandran, Deepak and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceeding of the ACL 2002*, pages 41–47.

Sekine, Satoshi. 2005. Automatic paraphrase discovery based on context and keywords between ne pairs. In *Proceedings of IWP 2005*.

Szpektor, Idan, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP 2004*, pages 41–48.

Turmo, Jordi, Alicia Ageno, and Neus Català. 2006. Adaptive information extraction. *ACM Computing Surveys*, 38(2):1–47.

Wu, Fei, Raphael Hoffmann, and Daniel S. Weld. 2008. Information extraction from wikipedia: Moving down the long tail. In *Proceedings of the KDD 2008. ACM*, pages 731–739.

Wu, Fei and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the CIKM 2007. ACM*, pages 41–50.

Zhao, Shiqi, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of the ACL 2008: HLT*, pages 780–788.

Zhao, Shiqi, Haifeng Wang, Ting Liu, and Sheng Li. 2009. Extracting paraphrase patterns from bilingual parallel corpora. *Natural Language Engineering*, 15(4):503–526.