Meaningful results for Information Retrieval in the MEANING project

Piek Vossen
Irion Technologies
Delftechpark 26
2628XH Delft, Netherlands,
Piek.Vossen@irion.nl

German Rigau, Iñaki Alegria, Eneko Agirre
IXA group
Manuel de Lardizábal, 1
20018, San Sebastián. Spain
rigau@si.ehu.es i.alegria@si.ehu.es
e.agirre@ehu.es

David Farwell
TALP Research group
Jordi Girona Salgado, 1-3
08034 Barcelona, Spain
farwell@talp.upc.edu

Manuel Fuentes
Agencia EFE
Espronceda, 32
28003-Madrid, Spain
mfuentes@agenciaefe.com

Abstract

The goal of the MEANING project (IST-2001-34460) is to develop tools for the automatic acquisition of lexical knowledge that will help Word Sense Disambiguation (WSD). The acquired lexical knowledge from various sources and various languages is stored in the Multilingual Central Repository (MCR) (Atserias et al 04), which is based on the design of the EuroWordNet database. The MCR holds wordnets in various languages (English, Spanish, Italian, Catalan and Basque), which are interconnected via an Inter-Lingual-Index (ILI). In addition, the MCR holds a number of ontologies and domain labels related to all concepts. During the MEANING project, the MCR has been enriched in various cycles. This paper describes the integration and evaluation of the MCR in a commercial classification and (cross-lingual) information retrieval system, developed by Irion Technologies. We carried out a series of task-based evaluations on English and Spanish news collections, for which indexes were built with and without the results of MEANING. The evaluations show that both recall and precision are significantly higher when using the enriched semantic networks in combination with WSD.

Introduction

The usefulness of wordnets for Information Retrieval and for Document Classification is not commonly accepted. Important evidence for this belief is a study of Voorhees (1994) that showed a decrease in scores for a wordnet-based approach in Trec-5. She claimed that linguistic techniques are only useful if they perform close to perfect. She also states that statistic techniques approximate NLP techniques by exploiting statistical correlations. A similar statement is made by Sanderson (1994) who suggests that wordnet-based approaches are only useful for retrieval if 90% or higher accuracy is achieved to detect the appropriate sense. This study was done by introducing artificial ambiguity in documents by substituting randomly chosen word pairs, e.g. *banana* and *kalashnikov*, with artificially ambiguous terms, e.g. *banana/kalashnikov*.

We however still believe that there is an enormous potential for wordnet-based approaches (also see Gonzalo et al 1998). In this paper we give evidence that wordnets can be exploited for generic information retrieval and classification tasks. The reason why this has not been evident is that the incorporation of wordnets is not trivial. Wordnets need to be used and integrated in a proper way to benefit from their

richness. The MEANING [1] project (IST 2001-34460; Rigau et al. 2002; 2003) tried to achieve this by pursuing the following goals:

- to enrich wordnets with more knowledge that is automatically acquired from corpora and the WWW;
- to improve Word Sense Disambiguation (WSD) using novel techniques in combination with the acquired knowledge;
- to develop a rich conceptual representation of text that is based on combinations of synsets associated with linguistic phrases;

The MEANING project developed tools for the automatic acquisition of lexical knowledge that will help WSD. The acquired lexical knowledge from various sources and various languages is stored in the Multilingual Central Repository (MCR), which is based on the design of the EuroWordNet database (Vossen, 1998). The MCR holds wordnets in various languages (English, Spanish, Italian, Catalan and Basque), which are interconnected via an Inter-Lingual-Index (ILI). In addition, the MCR holds a number of ontologies and domain labels related to all concepts. MEANING uses WordNet1.6 as an ILI to share lexical knowledge stored for each separate wordnet. During the MEANING project, the MCR has been enriched in various cycles.

This paper describes the integration and evaluation of the MCR data in a classification and (cross-lingual) information retrieval system, developed by Irion Technologies. In these applications, text is represented in the form of combinations of concepts that co-occur in linguistic phrases and where concepts are based on the synsets in the WordNet taken from the MCR. In a sense, the complete phrase represents a complex concept as whole, built up of interrelated sub-concepts consisting of synsets. Similarly, a query is considered as a phrase, representing one or more concepts. A query consisting of multiple concepts is then compared to phrases with multiple concepts. We carried out a series of task-based evaluations on English and Spanish news collections. The evaluation involved indexes built with and without the results of MEANING. The evaluation shows that both recall and precision are significantly higher when using the enriched semantic networks in combination with WSD.

The paper is structured as follows. In the next section, we briefly explain the conceptual indexing technology developed at Irion Technologies. Section 2 describes how the results of MEANING have been integrated in the Irion system. The following sections describe the evaluation. In Section 3 we describe the classification evaluation that was carried out on the Reuters news collection for English. Section 4 describes a cross-lingual retrieval evaluation on the same Reuters data and Section 5 another cross-lingual retrieval experiment on a database of news pictures with Spanish and English captions from the Spanish publisher EFE. Whereas the Reuters retrieval system used a classical vector-space document ranking, the EFE version uses a novel way of ranking based on the conceptual phrase representation. The EFE database is also used in an end-user evaluation task. This is described in Section 6.

1 Conceptual indexing at Irion Technologies

Irion Technologies (Delft, The Netherlands) developed a conceptual indexing technology, called TwentyOne, that combines statistical and language-technology approaches. TwentyOne is a two step system, where first, the relevant documents are collected using state-of-the-art statistical engines, and secondly, the best matching phrases from the relevant documents are collected. The statistical core-engine of TwentyOne produces a relevance ranking of text, using a standard vector-space weighting. It ensures fast and robust retrieval. The language-technology then has two major roles:

1. Maximize the recall of the statistical engine so that any document is found regardless of the wording and regardless of the query word choice;
2. Maximize the precision by conceptually matching queries with phrases in the documents rather than complete documents;

[1] http://www.lsi.upc.edu/~nlp/meaning/

The conceptual index represents concepts at a phrase level, which are very loosely defined as NPs. Within a phrase, a range of concepts is given where each concept correlates with a word, a combination of words or a part of a word, for example:

- The phrase *human rights* will represent a single concept that is lexicalized as a whole. Likewise it is translated to Dutch and German as a single word, as *mensenrechten* and *Menschenrechten* respectively. Note that this concept can still have relations to other concepts such as the hypernym *right* (in a very specific meaning) and *human*.
- The phrase *animal party* will represent 2 concepts, the separate concepts *animal* and *party* that co-occur, and so does *party animal* albeit a different combination.
- The single word *profile-based* will also represent two concepts *profile* and *based* as a co-occurring combination.

A conceptual representation of a phrase thus consists of a co-occurring sequence of synsets that express a particular relation to each other.

For building up a conceptual representation of a phrase, the TwentyOne system heavily relies on a multilingual semantic network, similar to EuroWordNet and the MEANING MCR. It uses multiword lookup, compound decomposition and WSD to map words within a phrase to concepts. Queries (user-queries or textual documents) are analysed in the same way. The TwentyOne system then uses a range of factors to compare phrases in documents with query phrases:

1. number of matching concepts between the query and each phrase,
2. degree of fuzziness mismatch between the query word and the phrase,
3. degree of derivational mismatch, compounding, etc.,
4. whether or not a synonym is used,
5. whether or not the same language is used.

The effect is first that documents with phrases (NPs) that include most concepts are shown first and, second, that documents with the same number of concepts but with the most similar wording with the query are shown first. The contextual effect of the phrase match is very powerful, as we will see later.

Because words are mapped to concepts from a language-specific wordnet that is linked through the ILI to all the other wordnets, it is possible to calculate a conceptual score for queries in other languages than the index language. Hence, any index can be queried in any of the languages that connected to the ILI.

2 Integrating MEANING in the Irion system

The MEANING results have been integrated in the Irion system in two ways:

1. we replaced the proprietary multilingual semantic network of Irion by the MCR database,
2. we developed a WSD system based on WordNet domains (Magnini et al 2002) which are integrated into the MCR.

The data from the MCR could be easily imported in the TwentyOne system because both the proprietary database of Irion and the MCR are based on the model of EuroWordNet. Within the Irion database, we simply replaced the concepts by the WordNet1.6 concepts and imported the mapping of the vocabulary for each language to these concepts from the MCR. Whereas the proprietary database has wordnets for English, Dutch, German, Spanish, Italian and French, the MCR has wordnets for English, Spanish, Italian, Basque and Catalan.

Although other possibilities could be explored (the use of other ontologies such as SUMO or the EuroWordNet Top Concept and a large amount of explicit semantic relations also integrated into the MCR), WSD was done using only WordNet domains (version 1.1.1, Magnini et al 2002) from the MCR. The WSD system has been implemented as a a text classifier that is trained with the Spanish and English words associated with domain labels, e.g. all synonyms related to the domain “legal”, and assigns a domain tag to

the text. The WSD system first assigns domain labels to the article as a whole, based on the complete content: so-called *microworld* tags. Next, it also classifies the separate NPs within each article using a window of 10 NPs (4 to the left and 5 to right). This results in one or more so-called *nanoworld* tags for each NP. All domains scoring above 60% confidence are assigned to have sufficient recall. The disambiguation then consists of the following process for each word in the NP:

1. Are there word meanings with domain labels that match any of the nanoworld tags? If yes, these meanings are selected.
2. If no, are there word meanings with domain labels that match the microworld tags? If yes these meanings are selected.
3. If no all meanings are selected.

The concept reduction as a result of the disambiguation is very effective. For the data obtained from EFE (see Section 6 below), the microworld-based reduction is about 48% for Spanish and 57% for English. In the case of the nanoworlds, the reduction is even higher: 52% for Spanish and 65% for English. Most of these reductions (about 44%) however relate to the factotum words (Magnini et al 2002). Factotum words are words such "be", "begin", "person" that are not specific to a domain and often have a very strongly preferred generic meaning. This generic meaning is labelled in WordNet domains and can be used to restrict the meanings when there is no other specific domain that applies to these words.

For each of the experiments described below, we built 3 types of indexes:

1. **NP** Indexes with NPs but without using wordnets: i.e. traditional string-based indexes.
2. **FULL** Indexes using wordnets, but without WSD: i.e. full expansion to all possible synonyms and/or translations.
3. **WSD** Indexes using wordnets and using word sense disambiguation: i.e. expansion limited to synonyms and/or translations within the context of the relevant domains, if any.

An example of an extracted NP is the following:

<NP ID="22">toxic to nerve cells </NP>

In the case of the NP index, the words are indexed as they occur (after normalization). In the case of the, the FULL index, the words ""toxic", "nerve" and "cell" are expanded to all the synonyms for all their possible meanings. These synonyms will thus also get a pointer to this document and this NP. In the case of the WSD index, we only added synonyms for word meanings that fit the domains assigned to the document and the NP.

We thus expect that the first type of index (**NP**) gives high-precision but lower recall because we do not generate a mapping through synonyms. You cannot find any documents with wordings different from the query.[2] The second type of index (**FULL**) will have a very high recall, because any possible mapping and wording is generated. The precision may drop because we also generate a lot of noise through irrelevant expansions. The third index (**WSD**) index will have recall and precision rates in between the others.

3 Document classification on the Reuters news collection

The first experiment was carried out using a document classification system that can be trained with example documents with categories. For training, we used the Reuters collection with English news. The experiment was restricted to the 23,307 files from a single month August 1996. The Reuters collection comes with classification codes that are embedded in the XML structure. We used the 125 topic codes, which can be organized hierarchically. We did not consider the hierarchical relations in the evaluation and training and treated each code separately. Multiple codes can be assigned to a single document.

[2] This is especially the case for smaller two-word queries, which is more normal for search engines.

Table 1: Recall and precision for Classification

	TEXT		NP		FULL		WSD	
RECALL	131.6	67.8%	138.8	72.3%	175.5	75.6%	188.2	80.7%
PRECISION	136.6	70.4%	143.4	74.7%	152.9	65.9%	168.2	72.2%
COVERAGE	194	83.2%	192	82.4%	232	99.5%	233	100%
F-MEASURE		69.0		73.5		70.4		76.2

The classification system has various options for testing and evaluation. One of the options is that a random test set is extracted from a training set. We thus trained the classification system with 22,074 files and set aside a test collection of 233 files. We then constructed the following classifiers from the training data:

1. HTM: the plain text is only normalized,
2. NP: NPs are extracted from the text but concepts are not expanded with wordnets,
3. FULL: NPs are extracted and fully expanded with wordnets,
4. WSD: NPs are extracted and expanded after disambiguation.

For each classifier, the same test files are excluded from training. RECALL (ρ) is then defined as $\rho = \alpha / \tau$, where:

> α = is the number of correct classes assigned to a test file,
>
> τ = is the total number of test classes that are associated with a file.

PRECISION (π) is defined as $\pi = \alpha / (\alpha + \beta)$, where:

> α = is the number of correct classes assigned to a test file,
>
> β = is the number of wrong classes assigned to a test file.

COVERAGE is then used to indicate on how many files the classifier is giving results above the threshold. For the experiments, we used a threshold setting of 0.7 (this is an empirically derived value for adequate results). Results below this threshold are ignored and are thus excluded from the COVERAGE. F-MEASURE is the harmonic mean of precision and recall. Table 1 then gives the results for the classifiers built with the different conceptual indexes:

Remarkably, the highest precision is obtained with NP and the highest recall with WSD. We see that FULL expansion leads to an increase of recall and a decrease of precision, compared to TEXT and NP. This is what we would expect. We also see that the coverage increased, i.e. there are more files for which there are results above the threshold. NP expansion leads to a lower recall (-3.3%) than FULL expansion but remarkably a higher precision (+8.8%). Here we see the effect of just using noun phrase extraction. Coverage is lower than for FULL expansion. Finally, best results are obtained for the disambiguated indexes. Recall is up to 80% and precision is slightly lower than NP expansion. However, coverage is now 100%. Apparently, the disambiguation expansion leads to results for documents with words that did not occur in the training set. This can be seen as a positive effect, whereas the negative effect is limited. Concluding, we see that the disambiguated expansion can lead to an increase of 12% in recall, 17% in coverage and still 2% increase of precision with respect to the baseline (TEXT). The f-measure shows clearly the superiority of the WSD results, with three points even over the NP results.

Table 2: Cross-lingual retrieval results on the Reuters collection

	English original "police cell"			**English** paraphrase "detention cell"			**Dutch** "politie-cel			**German** "Polizei-zelle"			**French** "cellule de police"			**Italian** "cella della polizia			**Spanish** "celda de la policía"		
	Q	**R**	**%**	**Q**	**R**	**%**	**Q**	**R**	**%**	**Q**	**R**	**%**	**Q**	**R**	**%**	**Q**	**R**	**%**	**Q**	**R**	**%**
NP	96	76	79	96	24	25	96	8	8	96	8	8	95	10	11	94	4	4	96	4	4
FULL	96	61	64	96	28	29	96	35	36	96	38	40	95	42	44	94	20	21	96	18	19
WSD	96	68	71	96	30	31	96	34	35	96	30	31	95	36	38	94	17	18	96	15	16

4 Crosslingual retrieval on the Reuters data

The same Reuters indexes were also used for a cross-lingual retrieval experiment. The TwentyOne retrieval system has a benchmark environment that can extract NPs from the indexed documents and create queries, where we measure if the same document from which the NP is extracted is returned within the top-ranked documents. Note that this measurement does not tell you anything about the quality of the other results. It can thus only be seen as a crude way to measure the recall of the system.

We thus automatically extracted NP-based queries from the indexes. Next, we manually selected 96 queries with a head and a modifier, where the head noun exhibits a clear case of homonymy or synonymy. For example, the noun *cell* that has clearly different meanings when combined with *police cell*, *cell phone*, *battery cell*, etc. From the complete NPs, two word queries were extracted. Next the original queries were modified by replacing the modifier by another context word that is semantically related, sometimes with a similar disambiguating effect and sometimes more neutral. An example of this sort of modified query would be *detention cell* instead of *police cell*. This resulted in about 96 paraphrased queries in English. Next the original queries were translated into the other languages recognized by the system: Dutch, German, French, Spanish and Italian.

We then run separate tests on the 3 types of indexes: NP, FULL and WSD, with the original words as query, the paraphrased English words or the translations of the originals. The results are shown in Table 2, where the rows represent the different indexes and the columns the results for each set of queries: original words from the NPs, paraphrased English words and translations.

In table 2, each query result column has 3 sub-columns:

Q = number of queries
R = recall, the number of times that the document from which the query was extracted occurs in the top 10 results
% = proportional recall

When we look at the original words used as a query, we see the best result on the NP index. The FULL index can only generate more noise by the expansion compared to the original words. This has pushed good results out of the top 10. We see that the WSD index has a positive effect because the recall recovers with 7%. When we look at the paraphrased English queries, we see that the recall dramatically drops for the NP index. This shows that the type of query is important to demonstrate the need for a wordnet-type of expansion. We see here that the WSD index gives best results.

The cross-lingual results can be compared with the paraphrased results. Obviously, the NP indexes perform poorest because the words are not translated at all (i.e. there is no expansion). The FULL index now has better results than WSD. Apparently, the noise generated cross-linguistically by giving all possible translations has a less negative effect compared to missing certain translations due to WSD.

In the above experiment, we used the proprietary wordnet database from Irion and we did not yet implement the conceptual scoring function that re-ranks the relevant documents on the basis of the overlap of concepts between the NPs and the queries, combined with the closeness of expression. The ranking was here based on the traditional statistical relevance ranking. In the next section, we describe a further experiment with the MEANING wordnets and with the conceptual re-ranking.

5 Cross-lingual retrieval on the EFE data

For this experiment, we indexed part of a multilingual database of pictures, called Fototeca, that was provided by the Spanish news agency EFE. We received a collection of 29,511 XML records with captions and corresponding pictures (from EPA and AP). These captions have 50 words of text on average. The captions are manually enriched for monolingual and multilingual access. This collection can be used to find pictures using text queries on the captions. Most of the captions were Spanish (26,546), about 10% were in English (2,965).

Again, we built the 3 types of indexes NP, FULL and WSD. In this case, however, we used the MEANING MCR data, which enables us to use the latest results of MEANING as well as use other languages (Basque and Catalan) for querying. In the case of the NP index, we built indexes for 5 languages: English, Spanish, Catalan, Basque and Italian. Instead of translating the original English and Spanish words they were simply copied to the other indexes for English, Spanish, Catalan, Basque and Italian. For example, the Basque index did not contain Basque translations but the literal Spanish and English originals. No synonym expansion was applied for English and Spanish and no translation for the other languages.

For indexes FULL and WSD, the Spanish and English indexes were expanded to synonyms and translated to English (in case of Spanish), Spanish (in case of English), and to Basque, Italian and Catalan (from both English and Spanish) with wordnets from the MCR. In the case of index FULL, all the meanings of the words in the articles have been taken and have been expanded to all synonyms and/or translations. In the case of WSD, we first excluded unlikely meanings using the WSD system and expanded all the remaining queries. For all three indexes queries can be made in any of the 5 languages: Spanish, English, Basque, Italian and Catalan, while the system returns both English and Spanish articles as possible results.

The queries were extracted as described previously for Reuters (Section 5). In this case, we automatically extracted Spanish NPs (e.g. “Una colisión en cadena”) and manually selected 2-3 word queries (e.g. query “colisión en cadena”) showing ambiguity or synonymy. We verified that other meanings and/or synonyms also occurred in the index, for example for *estrella* (*star*) we checked to make sure that it was used in both an *astronomical object* reading as well as a *leading actor* reading. Similarly, with *figura* we found that it was used in various different readings including *body*, *form, figure, character* and *statue*. Finally, we also looked at the relevance of the words to the corresponding pictures. This resulted in about 105 queries based on the original expressions extracted from the captions. From these we created paraphrase queries by replacing each context word with a synonym. Finally, the original queries were translated by native-speakers into English, Catalan, Basque and Italian.

The results of launching the queries on the 3 indexes are listed in Table 3. The results per index are given in the rows (NP, FULL and WSD) and the columns represent the different query sets: original Spanish words, paraphrased Spanish queries and the translated queries. The sub-columns are the same as above for Reuters.

Table 3: Retrieval results for multi word queries

	Spanish original			Spanish paraphrase			English			Catalan			Basque			Italian		
	Q	R	%	Q	R	%	Q	R	%	Q	R	%	Q	R	%	Q	R	%
NP	105	99	94	94	14	15	105	2	2	105	31	3	104	1	1	105	3	3
p1		60	57		9	1		0	0		21	2		1	1		2	2
p2		30	29		5	5		1	1		8	8		0	0		1	1
p3		9	9		0	0		1	1		2	2		0	0		0	0
FULL	105	96	91	94	71	76	105	39	37	105	70	67	104	50	48	105	39	37
p1		55	52		38	4		16	15		44	42		27	26		19	18
p2		33	31		27	29		17	16		22	21		19	18		15	14
p3		8	8		6	6		6	6		4	4		4	4		5	5
WSD	105	97	92	94	61	65	105	39	37	105	68	65	104	46	44	105	32	30
p1		60	57		39	41		21	2		48	46		27	26		20	19
p2		31	3		18	19		13	12		16	15		15	14		6	6
p3		6	6		4	4		5	5		4	4		4	4		6	6

The rows are slightly different. Each index has a row for the total results and three more rows for the 1st, 2nd and 3rd position (p1, p2 and p3) in the result list. We marked the best scores for the totals and for the 1st position (p1). We did not list the other positions from the top 10 because all the results listed the correct match in the top 3 or outside the top 10. The ranking algorithm was changed with respect to the Reuters experiments. The relevant documents were re-ranked on the basis of the overlap of concepts between the query and the NPs in the documents, as explained in Section 2.

The first thing to be noticed is the high recall. The best results are for the original Spanish words on the NP index: 94%. This is inherent to the conceptual phrase search. The search engine will select NPs that include all the query concepts and give preference to NPs that closely match the query. When we do not use wordnets, as in NP, the most equal phrases are likely to show up first, especially since the queries have been derived from the NPs and there are not that many NPs with all the query words.

We also see that we hardly lose anything when we use wordnets. The fully expanded index (FULL) scores 91% and the disambiguated index (WSD) scores 92%. This is a major difference with respect to the results reported for the Reuters experiments. In Reuters, the retrieval was based on the page score and not on the conceptual phrase score. The conceptual phrase matching thus adds precision. So even if the wordnets add more possible hits and more noise, the fact that the closest wordings are preferred selects the most appropriate results. This is also clear when we look at the p1 positions. Here NP and WSD score equally well.

When we look at the queries where a synonymous word was used (the 2nd column group, Spanish paraphrase), we see that the index without wordnets (NP) drops to 15% but the FULL index only drops to 76% and the WSD index drops to 65%. This clearly shows the usefulness of wordnets for information retrieval. We also see that WSD apparently removed certain synonyms that are useful, hence the difference of 10% between FULL and WSD. This indicates that the WSD settings might have been too strict (50% of the concepts have been excluded).

On the other hand, if we look at the p1 scores, we see that WSD scores better than FULL. This means FULL generates more noise that is interfering with the correct results for the 1st position but the correct results apparently still end up in the top 10. This also implies that the total results for FULL can be worse than WSD if

the index is bigger. In a bigger index there is more competition and the noisy results will push correct results out of the top 10. The pattern that we see for the synonyms also shows up for the cross-lingual retrieval. FULL mostly has best results and WSD is very close but scores better for p1. NP has dramatically bad results.[3]

The 1st position results can be seen as a measurement of precision. The disambiguated index thus has a better precision than the fully expanded index. These results are confirmed in the end-user evaluation that is described in the next section.

6 End-user evaluation

6.1 The goal of the experiment

The end-user evaluation was performed in a real scenario provided by Spanish news agency EFE, using the Fototeca database, the database used by EFE to provide pictures that correspond to news articles. Within MEANING, we designed a complete end-user evaluation framework for this database following (Walker, *et al*. 1997). The design was validated in a pilot test with a single user. In this pilot test, the user was asked to perform a set of tasks with different systems in a limited time. Finally, the user was asked to fill a questionnaire. With this pilot test, we planned to check the appropriateness and correctness of the whole evaluation framework including the task design, the questionnaire, the three Irion systems, the log files, the number of end-users that would be needed, etc. As a result of the pilot test, we slightly revised the set-up.

For the end-user evaluation, we used the same three indexes of the EFE Fototeca collection that are described in Section 6:

- EFE_NP: no use of wordnets.
- EFE_FULL: wordnets with full expansion, no disambiguation
- EFE_WSD: wordnets with expansion after disambiguation.

[3] Catalan scores almost as well as the Spanish synonyms. This shows that the languages are closely related. The fact that both the wordnets are developed by the same group may also be a factor.

6.2 The end-user tasks

The end-user final evaluation was performed by three different users: a, b, c. Each end-user tested the three different systems: EFE_WSD, EFE_FULL, EFE_NP, which we have renamed here A, B and C respectively. Each end-user had to perform twenty-one different tasks organized in three test sets (1, 2, 3) having seven tasks each. Thus, each end-user performed a total number of twenty-one different tasks using three different systems. There is no repetition of a given combination of user, system or test set. The final evaluation schema was as following:

	End-users		
Test sets	a	b	c
1	A	B	C
2	B	C	A
3	C	A	B

This schema tries to neutralize undesired side effects related to the relative performance of the users (some users are better than others when locating pictures) and the inherent difficulty of the tasks (some tasks are more difficult than other). Furthermore, from the log files we only took into consideration the total number of actions performed by the three systems.

The total time allowed for performing each test set was twenty minutes. After finishing each test set, the end-user took an additional ten minutes to fill out a questionnaire.

Each test set was designed to be self-explanatory. The end-user was to search for a set of picture to accompany a set of articles they were writing using a system located at a particular web page which provided access to the EFE Fototeca database. For each task, the end-user was told that they were preparing a news article on a given TOPIC with a given CONTEXT and was then asked to locate a picture showing some GOAL to serve as a visual. This is exemplified by News Article 10:

News Article 10
TOPIC = TERRORISMO
CONTEXT = Sigue la violencia en Colombia y especialmente en Medellín.
GOAL = Un entierro en Medellín.

In the task of News Article 10, the end-user is required to locate a picture showing a funeral in Medellín (GOAL), given the continuing violence in Colombia and especially in Medellín (CONTEXT) related to TERRORISM (TOPIC).

We designed the 21 tasks so that it is difficult to locate the pictures by a regular textual Information Retrieval System (like EFE_NP). For example, there are no captions in the database matching both *entierro* (funeral) and *Medellín*. In fact, there are only two pictures with *sepelio* and *Medellín*, *sepelio* also used to express the concept funeral. Furthermore, *entierro* is more common (35 occurrences in the database) than *sepelio* (14 occurrences). That is to say, the most common words, as opposed to the less common words, were used in presenting the GOAL and CONTEXT of each task. Furthermore, some of the tasks (three in total) were designed to locate English captions instead of Spanish captions. Proper noun phrase were mostly excluded. We also keep track of the following information in each task:

News Article 10
QUERY = entierro medellín
TEXT = sepelio medellín
RESULT = **FH_1205173 20040524** and **FH_1205172 20040524**

CAPTION=

Terrorismo
TRI:JUSTICIA-INTERIOR-SUCESOS,TERRORISMO

CATEGORÍAS SUPLEMENTARIAS : JUSTICIA-INTERIOR-SUCESOS PALABRAS CLAVE : JUSTICE EXPLOSION DE UNA MALETA BOMBA EN UNA DISCOTECA DE LA ZONA BANANERA DE URABÁ EN LA QUE FALLECIERON AL MENOS CINCO PERSONAS Y 93 RESULTARON HERIDAS , PRESUNTAMENTE COLOCADA POR LAS FARC , COLOMBIA 2004 FUNERAL VICTIMAS SM COLOMBIA SEPELIO VÍCTIMA BOMBA APARTADÓ : BOG302 MEDELLIN(COLOMBIA) 24/ 05/ 04 .- En el cementerio de San Pedro de Medellín se realizó el sepelio de la niña de 4 años de edad , María Fernanda Ramírez , una de las 7 víctimas de la bomba detonada en un centro nocturno de Apartadó . EFE/EDWIN BUSTAMANTE

PICTURE=

RESULT = **FH_1205173 20040524**

RESULT = **FH_1205172 20040524**

It is still possible to obtain the above results in English or Spanish because of the concept-based Information Retrieval system (like EFE_FULL and EFE_WSD) because in the Multilingual Central Repository we already have the concept: <entierro, sepelio, enterramiento> which has an equivalent translation through the ILI to the English concept <burial>.
After being instructed, the end-user queries the Fototeca database for an appropriate photograph using the system we were testing. When the system returns an ordered list of snippets (showing only a part of the text) the user reviews the results in order to select the most appropriate caption. Once a caption is selected, the system shows the corresponding picture. If the image is appropriate, the end-user clicks on a button labelled "This is the right picture". If, on the other hand, the picture is not appropriate, the end-user clicks on a button labelled "This is the wrong picture".

Table 4: Summary result figures for the final end-user evaluation

Three end-users	NP	FULL	WSD
SEARCH	110	64	**56**
HIGHLIGHT	105	**55**	60
DISAP.	57	28	**27**
CONFIRMED	20	19	**24**
UNDEC.	3	6	**1**
TOTAL	295	172	**168**

When the user is not sure, he clicks on a button labelled "Not sure about this picture". We also informed the end-user that if he did not find an appropriate photo the first time, he could try modifying the query, adding, removing or changing words from the original query. He could also select more than one picture for each news article. However, the total time for locating the appropriate pictures for each test set (seven tasks) was only twenty minutes.

6.3 The end-users

Three end-users were requested to perform the final end-user evaluation. We monitored all the activities of all the users by means of log files. All the tests were performed at the central office building of EFE in Madrid. They carried out all the activities in their usual workplace (office, computer, monitor, table, etc.).

6.4 The results

Table 4 summarizes the data we obtained from the log files. We count the total number of interactions performed with each system by the three end-users (TOTAL), the total number of searches (SEARCH), the total number of captions highlighted to see the corresponding picture when reading the caption text (HIGHLIGHT), the total number of pictures discarded after selected (DISAP.), the total number of confirmed pictures after selected (CONFIRMED) and the total number of pictures where the end-user was undecided (UNDEC.). It should be noted that the end-users had the same total time to perform the 21 tasks.

With respect to the total number of searches (SEARCH), we can see that in order to carry out the 21 tasks, the end-users made almost twice as many queries while using a text-based IR system (NP with 110) in comparison to a concept-based IR systems (FULL with 64 and WSD with 56).In other words, the users effort during search was reduced by almost half when they used a concept-based IR system. In addition, the total number of searches was significantly better (12.5%) for the system using word sense disambiguation (WSD with 56) as opposed to the one without (FULL with 64).

Regarding the total number of highlighted captions (HIGHLIGHT), we can also see that the end-users selected almost twice as many captions when using a text-based IR system (NP with 105) than when using a concept-based IR system (FULL with 55 and WSD with 60). This is because the user obtained essentially half of the false positives with a concept-based system. While it appears that in this case the FULL system outperformed the WSD system, we will see later that this is a misleading conclusion.

With respect to the total number of pictures rejected (DISAP.), we can see that the end-users discarded twice as many pictures when using a text-based IR system (NP with 57) than using a concept-based IR system (FULL with 28 and WSD with 27). That is to say, the users looked at half as many irrelevant pictures in order to locate the 21 desired pictures using a concept-based IR system. The total number rejected using the system performing disambiguation (WSD with 27) was essentially the same as that for the system without disambiguation (FULL with 28).

In terms of the total number of pictures confirmed (CONFIRMED), the three systems had similar behaviours (NP with 20, FULL with

19 and WSD with 24). This means that even with a poor text-based system (NP), the end-users having enough time were able to locate almost a complete list of pictures appropriate to the tasks (20 out of 21). However, with much less time (as it is derived from the log files) and half of the search effort (SEARCH) and total number of interactions (TOTAL) the end-users were able to locate an even more extensive list of appropriate pictures using the concept-based system with disambiguation.

Furthermore, the total number of pictures about whose relevance the user was unclear (UNDEC) was reduced with the WSD system (only one) in comparison with the other two systems (FULL with 6 and NP with 3). This is due in part to an interesting hidden behaviour, namely, that the WSD system also provided a better ranking of the captions. In other words, relevant captions were ranked higher giving the user greater confidence in the initial choice of captions and pictures (more confirmed, less undecided).

In summary, it seems that for difficult tasks (with synonyms or cross-lingual retrieval), using a concept-based IR system with WSD results in half of the searching effort, more confirmations, half the false positives, half the rate of rejection, fewer undecided pictures and half the total number of interactions. Although the results are preliminary, there is strong evidence with the end-user evaluation together with the previous Reuters and EFE CLIR experiments for suggesting that we performed better IR and CLIR with the help of the Multilingual Central Repository and appropriate WSD technology.

Regarding the questionnaire, it is not surprising that the end-users, who tested different questions (of variable difficulty) using different systems (with different performances), provided conflicting responses in regard to their perception of the systems' behaviors. We will not present the details of their responses here except to mention that they preferred system A (EFE_WSD) for future use.

7 Conclusions

This paper has summarized the results of a number of evaluations of the MEANING Wordnet database. It describes some larger tests with queries in various languages using the TwentyOne Search and Classification engine of Irion Technologies and an end-user evaluation in a real-world scenario on two months of captions and pictures from the EFE Fototeca database.

The integration required the use of the Spanish, English, Catalan, Basque and Italian wordnets from the MCR. It also involved the use of WordNet domains exported from the MCR and integrated in the WSD system developed by Irion Technologies. The disambiguation resulted in the reduction of 50% of the concepts.

With respect to classification, we have seen that disambiguated expansion can lead to an increase of 12% in recall, 17% in coverage and still 2% increase of precision with respect to the baseline (TEXT). The F-measure increased by 7.2%. With respect to retrieval, we saw significant improvement in recall for paraphrased queries (5%) and translated queries (15%-30%) on the Reuters data when using the MCR (with and without WSD). However, we lost 8% (using WSD) and 15% (using full expansion) on queries literally extracted from the text.

For the EFE database, we modified the ranking so that the queries are matched with concept combination in phrases (NPs). We have seen that the phrase structure helps to exclude the noise generated by the expansion with wordnets. Literal queries only dropped 2% and 3% when using wordnets and WSD, whereas paraphrased queries gained 50% to 60% and translated queries 35% and higher. In addition, when we took the correctness of the first result as a measure of precision, the WSD approach performed best for all types of queries.

Finally, we also described the end-user evaluation framework and the results obtained, which have been carried out by a three different users testing three different systems. This experiment confirmed the results from the query-based experiments. The productivity of the

end-users doubled and there was a clear effect in precision for the WSD-based system.

A concept-based IR system with WSD appears to beneficiary in difficult tasks (with synonyms or cross-lingual retrieval). Using half the search effort, it results in more confirmed photographs, half the false positives, half the number of rejected pictures, fewer uncertain selections and half the total number of interactions. The bottom line is that IR and CLIR can be improved with the help of the Multilingual Central Repository and appropriate WSD technology.
This end-user evaluation showed that both WSD and FULL increase productivity when searching for pictures in the Fototeca database. However, WSD significantly outperforms the FULL because the first result is more often the correct result and, as a result, users can quickly and correctly finish their task without going through the full set of responses.

It is also clear from these findings that a phrasal representation of the concepts in wordnets is important in order to achieve good results. For the future, we therefore want to further explore the possibilities for extracting a more detailed representation of the conceptual relations expressed in phrases. The current system, for instance, does not distinguish between *animal party* and *party animal* or between *Internet service on Java* and Java Internet Services because it cannot detect the conceptual relation between the concepts. This would also require higher precision WSD and more inferencing and reasoning which will allow a question such as *Who are the parents of Ghandi?* to be answered by a phrase akin to *Ghandi is the son of* ….

Acknowledgements

We want to thank the reviewers for their valuable comments. This work has been partially supported by the European Commission (MEANING IST-2001-34460).

References

Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, Piek Vossen "The MEANING Multilingual Central Repository" In Proceedings of the Second International WordNet Conference-GWC 2004 pg. 23-30 January 2004, Brno, Czech Republic. ISBN 80-210-3302-9

Fellbaum, C. (ed) (1998) WordNet. An Electronic Lexical Database, The MIT Press.

Gonzalo, J., F. Verdejo, I. Chugur and J. Cigarrán (1998) *Indexing with WordNet synsets can improve text retrieval*. Proceedings of the ACL/COLING98 Workshop on Usage of WordNet for Natural Language Processing. Sanda Harabagiu ed.

Magnini, B. and G Cavagliá (2000) *Integrating subject field codes into wordnet*. Proceedings of the Second International Conference on Language Resources and Evaluation LREC'2000, Athens, Greece.

Rigau, G., B. Magnini, E. Agirre, P. Vossen and J. Carroll (2002) *Unsupervised word sense disambiguation rivaling supervised methods*. Proceedings of COLING Workshop, Taipei, Taiwan.

Rigau, G and E. Agirre and J. Atserias (2003) *The MEANING project*. Proceedings of the XIX Congreso de la Sociedad Espagna para el Procesamiento del Lenguaje Natural (SEPLN'03), Alacalte Henares, Spain.

Sanderson, M. (1994) *Word sense disambiguation and information retrieval*. Proceedings of 17th International Conference on Research and Development in Information Retrieval, 1994.

Voorhees E, M. (1994) *Query expansion using lexical semantic relations*: Proceedings of the 17th Annual International ACM_SIGIR Conference on Research and Development in Information Retrieval.

Vossen, P. (ed) (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht.

Vossen, P., E. Glaser, H. Van Zutphen, R. Steenwijk (2004) Validation of MEANING, WP8.1 Deliverable 8.1, MEANING, IST-2001-34460, Irion Technologies BV, Delft, The Netherlands.

Vossen Piek, German Rigau, Iñaki Alegria, Eneko Agirre, David Farwell, Manuel Fuentes (2005) Validation of MEANING2, MEANING, IST-2001-34460, Deliverable 8.2.

Walker, M., D. Litman, C. Kamm, and A. Abella. (1997) *PARADISE: A Framework for Evaluating Spoken Dialogue Agents*. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, ACL-97, Madrid, Spain.