Some Linguistic Methods of Improving of the Quality of Document Retrieval on the Internet

Alexander Gelbukh[1,2], Grigori Sidorov[1], and Yoel Ledo-Mezquita[1]

[1] Natural Language and Text Processing Laboratory,
Center for Research in Computer Science, National Polytechnic Institute,
Av. Juan Dios Batiz s/n, Zacatenco 07738, Mexico City, Mexico
{gelbukh, sidorov, ledo}@cic.ipn.mx, www.gelbukh.com

[2] Department of Computer Science and Engineering, Chung-Ang University,
221 Huksuk-Dong, DongJak-Ku, Seoul, 156-756, Korea

Abstract. One of the problems of e-Business is to find relevant documents for making correct decisions. The main problem of the Internet is the huge amount of documents that makes it difficult to find the relevant ones, hence the importance of the methods allowing for improving the quality of document retrieval. We discuss some linguistic problems of document retrieval on the Internet related to the following natural language phenomena: (1) morphological processes: e.g., *takes*, *took*, *taken* are grammar forms of *take*, (2) polysemy and homonymy: most words have several senses, e.g., *bank* is a financial institution, shore, bench, etc., (3) non-linearity of syntactic relations: in case of a query that contains word combinations, the words forming a word combination can be separated by other words in the documents. Some linguistic-based methods and strategies related to the discussed problems are proposed that improve the quality of document retrieval or show the necessity of application of linguistic methods.

1. Introduction

Conducting e-business in Internet assumes business communication (such as offers and requests of goods or services) between people who do not coincide in space and time (hence the need in using the Internet), using automatic means due to the huge amount of information (search in huge databases) and the time constraints (the one who first discovers an advantageous offer gets it). Hence the vital necessity in robust natural language understanding technology for e-business, given that most of business communication is conducted in natural language. In addition, due to standardization difficulties, there is a general consensus on that the language of communication of e-business software agents in Internet will eventually be plain English. In particular, the problem of word sense disambiguation discussed below in this paper is one of the central problems in both automatic (agent-driven) and manual (traditional) business communication in Internet. Consider an intentionally simplistic example: a human or software agent that looks in Internet for, say, flowers, risks buying musical instruments instead if it interprets incorrectly the word *viola* in the description of the offer.

In addition, any person involved in decision-making in e-Business knows that though Internet is a great source of information, it is very difficult to find the specific data one is interested in. Usually search engines find the documents on the basis of keywords occurring "near" each other in the document. Some specialized techniques are applied for ranking the documents from the more relevant for the user to less relevant, according to the system's estimation. However, these methods of documents ranking fail very often.

Usually, the documents are texts written in natural language. Thus, application of linguistic techniques allows for improving quality of document retrieval. We discuss some linguistic problems of document retrieval on the Internet related to the following natural language phenomena:

(1) Morphological processes: e.g., *takes*, *took*, *taken* are grammar forms of *take*. This causes the necessity of identification of such forms, and, thus, the development of systems for morphological analysis, which is to be applied during indexing and/or during text search.

(2) Polysemy and homonymy: most words have several senses, e.g., *bank* is a financial institute, shore, bench, etc. It is necessary to distinguish between such words that are represented by the same letter string but express mean different meaning; such a task is called automatic word sense disambiguation. Indeed, a user searching for *international banks in Mexico* should not be presented

with documents about an *international campus on the banks of the Usumacinta River in Mexico*. Thus, some elements of semantic analysis are necessary to improve the quality of text search.

(3) Non-linearity of syntactic relations: if the query contains a word combination, the words of this word combination can be separated by other words in the documents. For example, for the query *Mexican banks* the user expects to get the document with the phrase *Mexican hotels and banks*, but not those with the phrase *For any Mexican banks in Madrid are easy to find.*

The paper is organized as follows. First we discuss the problems related to automatic morphological analysis and present an approach that allows for rapid development of such systems for different languages with little effort. Then we evaluate the plausibility of application of word sense disambiguation in information retrieval. We show that information retrieval will benefit from application of word sense disambiguation methods. Finally, we compare two possibilities of searching for word combinations: a method based on syntactic analysis that determines important dependencies between words, and a bigram method that only considers immediately neighboring words. We show that the linguistically-motivated method based on syntactic analysis shows much better retrieval quality.

2. Automatic Morphological Analysis

Natural language morphology studies the structure of words and its relation with grammatical categories of a given language. The objective of the automatic morphologic analysis is to carry out morphologic classification of word forms. For example, the analysis of the grammar form *takes* shows that it is a verb in *present* tense, *3rd* person, *singular* number, and its normalized form (lemma) is (to) *take*.

In case of English, the problem of morphological analysis is not so difficult—usually the words have two or three grammar forms that differ from their lemma in a very simple and regular way, the only exception being the irregular verbs. In case of the majority of the other European languages, though, it is much more complicated. For example, Spanish has 65 main verb forms (e.g., *den* 'let them give'), or more than 200 cliticized forms (e.g., *dénmelo* 'let them give him to me'). The situation is even more complicated in Slavic languages, like Russian or Czech. Say, nouns in Russian have 12 forms, adjectives more than 30, and verbs more than 150 forms (taking into account the participles).

The systems of document retrieval should be able to identify different grammar forms as belonging to the same lemma to ensure completeness of the retrieval results. Thus, some elements of morphological analysis are necessary for systems in any language, even in the morphology-poor English language. In this section, we describe an approach that allows for construction of morphological analysis modules with little effort and time; detailed description of other approaches can found in [2, 6, 11].

The complexity of the morphological system of a language for the task of automatic analysis does not depend very much on the number of the grammar classes or on homonymy of flexions, because the algorithms for processing flexions are rather straightforward. Instead, it mainly depends on number of stem alternations that cannot be guessed without consulting a dictionary. For example, consider Spanish verbs *mover–muevo* 'to move / I move' vs. *correr–corro* 'to run / I run'; looking at the verbs, it is impossible to deduce whether they have alternations or not. The same problem is present in English, for example, looking at verbs *to take* vs. *to brake,* one cannot say which one is irregular; this is the dictionary knowledge.

At one extreme is the approach to morphological analysis that stores all grammatical forms in a dictionary, along with the lemma and all necessary grammatical information associated with each form. With this approach, a morphological system is just a very large several-column database. Modern computers have the possibility of storing databases containing all grammatical forms for large dictionaries of inflective languages: a rough approximation for Spanish is 20 to 50 megabytes. Nevertheless, these models have their own shortcuts: it is difficult to add new words to the dictionary, it is impossible to implement the processing of unknown words, etc.

Thus, it is necessary to develop linguistically-based morphological models and the corresponding algorithms. However, not all methods are equally convenient in use and easy to implement. There are two important points that make the differences between different approaches:

- Static or dynamic method of processing of stem allomorphs (variants of the stem, such as *goose*/*geese*), and
- Morphological models that are used.

There are two methods of processing of stem allomorphs: static and dynamic. With the static method, all stem allomorphs are stored in the dictionary (usually there are two to four allomorphs, so the dictionary

size is not significantly affected; note that the majority of words (say, in Russian more than 70 percent) do not have stem alternations. The allomorphs are generated beforehand; this is not difficult because the information about each stem is available.

With the dynamic method, allomorphs are constructed dynamically, with the aim of reconstructing allomorphs that would be stored in the dictionary with the static method. The corresponding rules cannot be standardized, so the number of such rules is quite large, usually more than a thousand rules. Besides, they do not have any intuitive correspondence in common knowledge of the language. For example, in order to generate the dictionary stem for Russian stem *okon-* '*window*', it is necessary to delete the inner *-o-*, which gives *okn-*. A corresponding example for English—just to demonstrate the kind of processes—is *took*: it is necessary to change *-oo-* to *-a-* and add *-e* to obtain *take*. Without any beforehand information about the possible type of the stem this is difficult and leads to the necessary to develop and apply many unintuitive rules. This method is of high complexity (so-called NP-complete) [4]. Therefore, the static method is more reasonable and easy to implement than the dynamic one.

The other dilemma deals with the kind of morphological models. The obvious direct way for developing the morphological models is to create a new morphological class for any paradigm that exists in the language, thus, the number of classes is calculated up to 1500 for Czech [8] or 1000 for Russian [1]. These classes are artificial, created for the purposes of analysis.

The other possibility is to use the morphological models that already exist for generation. Say, in case of Russian there are about 40 morphological classes. These models are described in traditional grammars, because these grammars are oriented to generation. Besides, they correspond very well to the intuition of native speakers. To use these models for analysis, it is necessary to apply special techniques that allow for applying generation instead of direct analysis. Usually, generation is much simpler than analysis. Such a technique is known in artificial intelligence as "analysis through generation." In our case, it is applied as follows: The system first generates all possible hypotheses based on the possible flexions and then tries to generate the grammar forms according to each hypothesis using the corresponding stem and its morphological class taken from the dictionary. Note that there is a small number of classes, while the peculiarities of words are described using grammar marks for words, such as the presence of alternations, the absence of singular (*pluralia tantum*), etc. These marks are interpreted during the process of analysis/generation.

Obviously, it is much easier for development of a system to have a small number of morphological classes, which correspond very well to the intuition of speakers. Sometimes these classes already exist, but if they do not exist, it is easier to characterize the words in a given language applying the simple and intuitive classification.

Accordingly, we suggest an approach based on static method of processing of stem allomorphs (i.e., all allomorphs are stored in the dictionary) and on application of the natural morphological models created for generation using the analysis-through-generation procedure.

The first stage of our method is data preparation. The words of a language are described in terms of the chosen morphological models. Then the stem dictionary is generated with all possible allomorphs of each stem. Note that the stem allomorphs should be marked according to the algorithm of their generation—for example, first, second, etc. This information is necessary during grammar form generation, namely, for choosing the correct stem allomorph.

The next stage is the development of the algorithm of morphological analysis. The following modules (parts of algorithm) are necessary:

1. Module of generation of hypotheses. These hypotheses are based on the correspondence between flexions and sets of possible values of grammar categories (flexion → values), e.g., in English, flexion *-s* can express plural for nouns or 3rd person singular for verbs, etc.
2. Module of choice of stem allomorphs. This choice is based on the correspondence between the sets of values of grammar categories of morphological classes and the number of the stem allomorph (values → number of the stem allomorph). For example, consider English verb stems *verify/verifi-* as allomorphs; the first allomorph is used for the present tense (except for the 3rd person singular form), and the second one for the past or present 3rd person singular form. This can be indicated using masks, patterns, direct programming, etc. Note that we do not need the reverse correspondence because we apply this module only for generation.

3. Module of choice of flexions: which flexion is used for a given set of grammar categories of a given class (values → flexion). For example, in English for plural of nouns the flexions *-s* or *-es* are used depending on the last letters of the stem.

The irregular forms should be processed separately. They are stored in the dictionary with their lemma and values of grammar categories (number, tense, etc.). Therefore, their analysis is just searching in the dictionary (the hypothesis of the irregular form with zero flexion should always be considered). Their generation also consists of searching in the dictionary, though for the lemma and the corresponding values of grammar categories.

The procedure of generation is very simple. The input is (1) a set of values of grammar categories and (2) a string that identifies the word (stem allomorph or lemma). The procedure implies obtaining the information from the dictionary (the morphological class, etc.), choice of the correct stem allomorph, and choice of the correct flexion (see the above description of the corresponding modules).

With the suggested approach, the procedure of analysis is simple. The input is a string of characters. The procedure is as follows: (1) formulation of the hypotheses for all possible flexions, (2) a call of the generation procedure for each hypothesis, (3) comparison of the result of generation with the input. If they coincide then the hypothesis is correct. Note that it is important to apply generation because otherwise some incorrect forms would be analyzed as correct ones, for example, **taked* (instead of *took*). Indeed, both the stem *take-* and the flexion *-d* exist, but they are incompatible, which is verified through generation (the correct form *took* will be generated, which does not coincide with the input **taked*). Not analyze the words incorrectly is important not to generate wrong analyses. For example, the string *learned* stands only for past tense of the verb *to learn*, but not for its past participle, which cannot be verified without generation and comparison (cf. *lesson learnt* vs. *question asked*).

We have developed the systems for automatic morphological analysis on the basis of this approach for Russian, English, and Spanish languages. It took relatively little time and effort: from several weeks of work of one person for English, several months for Spanish [2] and up to a year for Russian [9], which has a more complex morphology.

3. Automatic Word Sense Disambiguation

Words in a typical explanatory dictionary have different meanings (senses); this phenomenon is known as *polysemy*. However, a word has only one of these dictionary senses in each specific text. The problem of choice of a word sense used in a given text is called a word sense disambiguation problem. There are different methods of word sense disambiguation. They can be classified into two main groups: statistical methods (see, for example, [7]) and methods based on application of different knowledge sources (see, for example, [12]).

Below we describe an experiment that allows for evaluation of the usefulness of using word sense disambiguation techniques in Internet search. Indeed, if texts are indexed by word senses instead of just letter strings, then given a query $bill_1$ 'financial document', the system would not present the documents containing $bill_2$ 'garden instrument'. How often does this situation occur? The idea behind our evaluation is as follows: If word senses are too close (too similar), then, on the one hand, the user will be unable to distinguish them for his or her informational need and, on the other hand, automatic word sense disambiguation in the texts of the documents will not be reliable. Therefore, an information retrieval system should not distinguish between such senses, not to compromising its recall (the completeness of the search results). However, if the senses are sufficiently different, then an information retrieval system should distinguish between them to improve its precision (absence of noise in the results) because they represent different meanings, some of them being related to the user information need and others being unrelated to it.

We propose to measure the semantic distances (similarity) between different senses of the same word. Note that such measurement normally is useful for specific purposes because usually the distance is measured between words. We measure the distance between two given senses as the relative number of equal or synonymous words in their definitions in an explanatory dictionary (in our experiments with Spanish texts we used the Anaya explanatory dictionary of Spanish). By equal words we mean equal lemmas, i.e., we do not distinguish words in different morphological forms (*go*, *goes*, *gone*, etc.). For detecting synonyms we use a Spanish dictionary of synonyms.

We show that a considerable part of word senses of the same word are quite different, and, thus, they should be distinguished for the purposes of information retrieval. On the other hand, another considerable

part of word senses of the same word are rather close; these senses either should not be distinguished in information retrieval or their definitions should be changed.

We used Anaya dictionary as a source for definitions of the word senses. This dictionary has more than 30,000 headwords, which have more that 60,000 word senses. For morphological processing, we applied a Spanish morphological analyzer and generator developed in our laboratory [2]. All definitions in the dictionary were normalized and the part of speech tagging procedure was applied to determine the part of speech of each word [10]. We also used a synonym dictionary of Spanish that contains about 20,000 headwords. This dictionary is applied for detecting synonymous words in definitions when we measure the distances between senses.

Below we give an example of the data (normalized definitions from Anaya dictionary) that was used. The definition of the Spanish word *fabricar* '*fabricate'* in one of the senses is as follows:

Fabricar: *transformar materias primas en productos semielaborados, o estos en productos ya acabados.*

'*Fabricate:* *transform raw materials into semi-elaborated products, or the latter into already terminated products.*'

The normalized version of this definition is as follows:

Fabricar: $transformar_{\text{verb}}$ $materia_{\text{noun}}$ $prima_{\text{adj}}$ en_{prep} $producto_{\text{noun}}$ $semielaborado_{\text{part}}$ $,_{\text{punct}}$ o_{conj} $esto_{\text{pron}}$ en_{prep} $producto_{\text{noun}}$ ya_{adv} $acabado_{\text{part}}$ $._{punct}$

'*Fabricate:* $transform_{\text{verb}}$ raw_{adj} $material_{\text{noun}}$ $into_{\text{prep}}$ $semi\text{-}elaborated_{\text{part}}$ $product_{\text{noun}}$ $,_{\text{punct}}$ or_{conj} the_{det} $latter_{\text{adj}}$ $into_{\text{prep}}$ $already_{\text{adv}}$ $terminated_{\text{part}}$ $product_{\text{noun}}$ $._{\text{punc}}$'

where *conj* stands for conjunction, *adv* for adverb, *prep* for preposition, *adj* for adjective, *det* for determiner, *part* for participle, and *punct* for punctuation mark. There are some words that were not recognized by our morphological analyzer (about 3%), which are marked as *unknown*.

It is obvious that at the stage of comparison it is desirable to ignore the auxiliary words (such as conjunctions or articles) because normally they do not add any semantic information or add some arbitrary information; for example, prepositions often depend on the subcategorization properties of a particular verb in the given language.

In the experiment, we measured the similarity between two different word senses of the same word. We used the commonly accepted measure of similarity between two texts that is analogous to the well-known Dice coefficient, see, for example [5]:

$$S(t_1,t_2)=\frac{|W_1\cap W_2|+|W_1\circ W_2|}{\max(|W_1|,|W_2|)}$$

where W_1 and W_2 are sets of words in the two definitions; $|W_1\cap W_2|$ stands for the number of words that occur in both texts, while the symbol "$\circ$" indicates that we calculate such intersection using synonyms.

The processing procedure was as follows. For each word in the dictionary, we measured the similarity between its senses. Obviously, words with only one sense were ignored. In Anaya dictionary, there are about 13,000 words with only one sense (out of 30,000 words in total). Since similarity is a symmetrical relation, it was calculated only once for each pair of senses of each word. Note that we considered homonyms as different words and not as different senses. Normally, this is the way that they are represented in the dictionaries: as different groups of senses. Besides, the homonyms by definition have different meanings, so this mode of their treatment is the most appropriate for our research.

We conducted several experiments with different weighting coefficients, i.e., different manners of treating synonyms. The results show that the weighting of synonyms obviously changes the obtained values, decreasing the similarity.

Since the similarity is a fraction, some fractions are more probable than others; specifically, due to a high number of short definitions, there were many figures with small denominators, such as 1/2, 1/3, 2/3, etc. To smooth this effect, we represented the experimental results by intervals of values and not by specific values. The number of intervals can be different, but should not be very large. We use five intervals, where the first interval contains the number of sense pairs with zero similarity. We use this interval because we know beforehand that there are many senses without intersection. The other intervals are equal, see Table 1.

Table 1. Sense pairs per intervals in the Anaya dictionary.

Interval of distances	Number of sense pairs	Percent of sense pairs
0.00-0.01	46205	67.43
0.01-0.25	14725	21.49
0.25-0.50	6655	9.71
0.50-0.75	600	0.88
0.75-1.00	336	0.49

Our experiment presents an argument for (1) the possibility of improvement of information retrieval by word sense disambiguation because overwhelming majority of word senses are significantly different and (2) for the need to merge some sense pairs into larger senses useful for information retrieval, or maybe for the need to change their definitions in the dictionary.

4. Automatic Detection of Word Combinations

One of the advanced options for information retrieval is search using word combinations, such as *to polish leather*. A traditional search engine will present the documents that contain both these words located not very far one from another. However, this is not a guarantee that the words are related; their co-occurrence can be mere coincidence, for example, *to polish stone using leather.* To guarantee that the documents are relevant for the query expressed by the word combination, it is necessary to perform syntactic analysis and detect that there is a syntactic relation between the two words. To what degree can it improve the performance of the search engine? Below we present an evaluation of the use of syntactic relations as compared with the bigram methods, which merely consider co-occurrences of the words.

We experimented with a randomly selected Spanish text downloaded from Internet (*Cervantes Digital Library*). In our experiment, we used a probabilistic parser and a context-free grammar with unification for Spanish language developed in our Laboratory [3]. Namely, we applied a program that (1) performs syntactic analysis, (2) obtains word combinations with corresponding relation between words, (3) filters them, and (4) stores them in a database.

For filtering, we use both syntactic and morphological features. Say, we filter out the relations with pronouns and articles according to the morphological filters. In addition, we apply syntactic filters, according to which only word combinations that have the following syntactic relations are stored in the database: verb–subject, verb–object (direct or indirect), noun–modifier (an adjective or another noun), verb–modifier (adverb). All other syntactic relations are discarded. The name of relation is stored as well.

Some special cases are: (1) coordinative relation (for example, *to read newspaper and magazine* should give two word combinations *to read newspaper* and *to read magazine*), so, we split such a phrase; (2) relation with a preposition. In case of prepositions, we took the dependent word of the preposition and marked its relation with the governing word of the preposition (the head of the prepositional phrase). This is justified by the fact that prepositions usually express grammar relations (say, in some languages these relations can be expressed by grammatical cases), so they are not important for lexical links. On the other hand, the choice of a preposition is important linguistic information. Therefore, in this case we store all three members.

As a baseline we used a method of gathering word combinations that takes all word pairs that are immediate neighbors (bigrams). We incorporated certain intelligence into the baseline method. Namely, after the modification, it ignores the articles and takes into account the prepositions. We will consider an example of our analysis for the Spanish sentence

Mamá compró una torta pequeñita y un pastel con una bailarina con zapatillas de punta.
'*Mother bought a little bun and a cake with a dancer in ballet-shoes.*'

The following syntactic tree corresponds to this sentence. Indentation represents dependencies in the tree: for example, the verb in the line 1 has dependents in the lines 2, 14, and 15; the conjunction in the line 2 has dependents in the line 3 and in the line 6, etc. Note that the words are normalizes morphologically, as described above. We mark with boldface the syntactic categories used in our grammar: ***V*** stands for verb, ***N*** for noun, ***SG*** for singular, etc. The name of syntactic relation is indicated in curly brackets. Note that the name of the relation is stored with the dependent word, because the governor can have several

dependents. In parenthesis given are the word and its lemma along with their English translation, e.g. (*compró: comprar* / *bought: to buy*).

```
1  V(SG,3PRS,MEAN) (compró: comprar / bought: to buy)
2  ├─CONJ_C {obj}  (y: y / and: and)
3  │  ├─N(SG,FEM) {coord_conj} (torta: torta / bun: bun)
4  │  │  └─ADJ(SG,FEM) {mod}  (pequeñita: pequeñito / little: little)
5  │  │    └─ART(SG,FEM) {det} (una: un / a: a)
6  │  └─N(SG,MASC) {coord_conj}  (pastel: pastel / cake: cake)
7  │    └─PR {prep}  (con: con / with: with)
8  │      └─N(SG,FEM) {prep} (bailarina: bailarina / dancer: dancer)
9  │        ├─PR {prep}  (con: con / in: in)
10 │        │ └─N(PL,FEM) {prep}  (zapatillas: zapatilla / shoes: shoe)
11 │        │   └─PR {prep}  (de: de / of: of)
12 │        │     └─N(SG,FEM) {prep} (punta: punta / point: point)
13 │        └─ART(SG,FEM) {det} (una: un / a: a)
14 ├─N(SG,FEM) {subj} (mamá: mamá / mother: mother)
15 └─$PERIOD  (.: .,)
```

The following word combinations were found in this sentence. Note that the word combinations 4 and 7 are filtered out due to the morphological filters.

1. *comprar* (obj) *torta*{Sg} (*buy* (obj) *bun*{Sg})
2. *comprar* (obj) *pastel* {Sg} (*buy* (obj) *cake* {Sg})
3. *torta* (mod) *pequeñito* (*bun* (mod) *little*)
4. * *torta* (det) *un* (*bun* (det) *a*)
5. *pastel* (mod) [*con*] *bailarina* {Sg} (*cake* (mod) [*with*] *dancer* {Sg})
6. *bailarina* (mod) [*con*] *zapatilla* {Pl} (*dancer* (mod) [*with*] *shoe* {Pl})
7. * *bailarina* (det) *un* (*dancer* (det) *a*)
8. *zapatilla* (mod) [*de*] *punta* {Sg} (*shoe* (mod) [*with*] *point* {Sg} *//= ballet shoe*)
9. *comprar* (subj) *mamá* {Sg} (*buy* (subj) *mother* {Sg})

The parsed text contains 741 words grouped in 60 sentences. The average length of a sentence was 12.4 words. Apart, we marked syntactic relations in these sentences manually. For the baseline method, the total number of words was 588 because among 741 words there were 153 articles and prepositions in the sentences. The following results were obtained. The total number of correct word combinations (determined by the manual markup) was 208. Of these, 148 word combinations were found by our method. At the same time, the baseline method found correctly as few as 111 word combinations. On the other hand, our method reported incorrectly only 63 word combinations (false alarms), while the baseline method marked as possible word combinations $588 \times 2 - 1 = 1175$ pairs, of which $1175 - 111 = 1064$ are false alarms.

These figures give us the following values of precision and recall (precision is the share of the correctly reported items among all reported items; recall is the share of reported items among the items that ideally should be reported). For our method, the precision was $148 / (148 + 63) = 0.70$ and recall $148 / 208 = 0.71$. For the baseline method, the precision was $111 / 1175 = 0.09$ and recall $111 / 208 = 0.53$. It can be seen that the recall of our method is better and precision is much better than those of the baseline method.

5. Conclusions

We have analyzed the plausibility of application of three linguistically-based techniques for document retrieval in Internet.

The first problem is related with identification of different grammar forms as "the same word," which is a necessary stage of information retrieval process. It is performed using a system of automatic morphological analysis. We suggested an approach that allows for development of such systems with little time and effort for different languages. The approach is based on static processing of allomorphs and the analysis-through-generation methodology.

The second problem touched upon distinguishing of different senses of the words in information retrieval. We have shown that the distinguishing of words senses allows for better satisfaction of the user's information need. However, not all senses are to be distinguished: some senses are too similar to be reliably distinguished both by the user and by the system; these senses are to be merged into larger ones.

The third problem was related to advanced retrieval strategy using word combinations. We have described an experiment that demonstrates that application of syntactic analysis increases recall and precision of the search engine in case of queries containing word combinations.

Acknowledgements

The work was done under partial support of Mexican Government (CONACyT, SNI), IPN (CGPI, COFAA, PIFI), and Korean Government (KIPA Distinguished Visiting Professorship program). The first author is currently on Sabbatical leave at Chung-Ang University. We thank Prof. Igor A. Bolshakov for useful discussions.

References

1. Gel'bukh, A.F. Effective implementation of morphology model for an inflectional natural language. *J. Automatic Documentation and Mathematical Linguistics*, Allerton Press, vol. 26, N 1, 1992, pp. 22–31.
2. Gelbukh, A. and G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: *Computational Linguistics and Intelligent Text Processing*. Proc. CICLing-2003, 4th International Conference on Intelligent Text Processing and Computational Linguistics. *Lecture Notes in Computer Science,* N 2588:, Springer-Verlag, pp. 215–220.
3. Gelbukh, A., G. Sidorov, S. Galicia Haro, I. Bolshakov. Environment for Development of a Natural Language Syntactic Analyzer. In: *Acta Academia 2002*, Moldova, 2002, pp. 206–213.
4. Hausser, R. Three Principled Methods of Automatic Word Form Recognition. Proc. of *VEXTAL: Venecia per il Tratamento Automatico delle Lingue*. Venice, Italy, 1999. pp. 91–100.
5. Jiang, J.J. and D.W. Conrad. From object comparison to semantic similarity. In: *Proc. of Pacling-99* (Pacific association for computational linguistics), August, 1999, Waterloo, Canada, pp. 256–263.
6. Koskenniemi, K. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD Thesis. University of Helsinki, 1983. 160 p.
7. Manning, C. D. and Shutze, H. *Foundations of statistical natural language processing*. Cambridge, MA, The MIT press, 1999, 680 p.
8. Sedlacek, R. and P. Smrz, A new Czech morphological analyzer AJKA. Proc. of *TSD-2001. Lecture Notes in Computer Science*, N 2166, Springer-Verlag, 2001, pp. 100–107.
9. Sidorov, G. O. Lemmatization in automatized system for compilation of personal style dictionaries of literature writers (in Russian). In: *Word by Dostoyevsky* (in Russian), Moscow, Russia, Russian Academy of Sciences, 1996, pp. 266–300.
10. Sidorov, G. and A. Gelbukh. Word sense disambiguation in a Spanish explanatory dictionary. Proc. *TALN-2001*, Tours, France, July 2–5, 2001, pp. 398–402.
11. Sproat, R. *Morphology and computation*. Cambridge, MA, MIT Press, 1992, 313 p.
12. Wilks, Y. and M. Stevenson. Combining weak knowledge sources for sense disambiguation. Proc. *IJCAI-99*, 1999, 884–889.