Serdica J. Computing **3** (2009), 133–158

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

AUTOMATIC IDENTIFICATION OF FALSE FRIENDS IN PARALLEL CORPORA: STATISTICAL AND SEMANTIC APPROACH

Svetlin Nakov

Abstract. False friends are pairs of words in two languages that are perceived as similar but have different meanings. We present an improved algorithm for acquiring false friends from sentence-level aligned parallel corpus based on statistical observations of words occurrences and co-occurrences in the parallel sentences. The results are compared with an entirely semantic measure for cross-lingual similarity between words based on using the Web as a corpus through analyzing the words' local contexts extracted from the text snippets returned by searching in Google. The statistical and semantic measures are further combined into an improved algorithm for identification of false friends that achieves almost twice better results than previously known algorithms. The evaluation is performed for identifying cognates between Bulgarian and Russian but the proposed methods could be adopted for other language pairs for which parallel corpora and bilingual glossaries are available.

ACM Computing Classification System (1998): H.3.3, I.2.7.

Key words: Cognates, false friends, identification of false friends, parallel corpus, cross-lingual semantic similarity, Web as a corpus.

1. Introduction. Words in two languages that have orthographic or phonetic similarity are often perceived as similar by meaning but such perception sometimes could be wrong. Depending of their meanings such pairs of words could be classified as cognates, partial cognates or false friends.

Cognates are pairs of words in different languages that have similar spelling and similar meanings. *Partial cognates* are pairs of words that have similar spelling and could have the same meaning in some contexts but different meanings in other contexts. *False friends* are pairs of words in different languages that have similar spelling and are perceived as similar but have different meanings.

There is a little confusion about the term *cognates* in the classical linguistics and in the computational linguistics. In the classical linguistics *cognates* means words in related languages with common origin which sometimes have similar spelling but not always. For example the Bulgarian words *роза* [roza] and *гюл* [gyul] (both meaning *rose*) have developed from the same ancestor Old Persian word **vrda-* but are entirely different in spelling. Computational linguists like [35] and [18] ignore the origin of the words and define cognates as pairs of words of different languages which share "obvious" phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. Following their definition for the rest of this paper we will use the term *cognates* to denote words that have similar spelling and meaning.

Example of cognates are the words *слънце* [sləntse] in Bulgarian and *солнце* [solntse] in Russian both meaning *sun*. As an example of partial cognates we have the Bulgarian word *син* [sin] and the Russian word *синий* [sini:] These words have similar spelling (only different inflection) and share the common meaning of *blue* but the Bulgarian *син* has one more commonly used meaning: *son*. False friends are for example the Bulgarian word *бистрота* [bistrota] and the Russian word *быстрота* [bəistrota] meaning respectively *clearness* in Bulgarian and *quickness* in Russian.

False friends are not only important when studying foreign languages, but have application in various natural language processing tasks like word alignment, statistical machine translation, word sense disambiguation, automated quality control for translators and others.

Our objectives are to design and evaluate an unsupervised algorithm that automatically extracts pairs of false friends from given parallel corpus aligned at sentence level. We don't want to distinguish between cognates and partial cognates so we are interested in finding only false friends – words perceived as similar and having different meanings in all contexts.

Our experiments are particularly focused on identification of false friends between Bulgarian and Russian, but the methods we describe are applicable to other language pairs as well.

Bulgarian and Russian are highly inflectional languages so cognates and false friends can appear between different parts of speech and different wordforms. Our objective is to indentify pairs of false friends including different part of speech and different word forms. For example we are interested in finding false friends like the Russian *могла* [mogla] (*she were able to do something*) and Bulgarian *мъгла* [məgla] (*fog*). Another example is the Russian *копейки* [kopeiki] (*cents*) and the Bulgarian *къпейки* [kəpeiki] (*during the act of taking a bath*).

The remaining of the paper is organized as follows: Section 2 discusses previous research in the area of automatic identification of cognates and false friends from text corpora. It includes methods for measuring orthographic similarity, methods for identifying false friends from parallel corpora and methods for measuring cross-lingual semantic similarity. Section 3 describes our algorithms for identifying cognates and false friends. It describes our approach to identifying candidates for cognates and false friends, our statistical method for extraction of false friends in parallel corpora and our method for extracting cross-lingual semantic similarity from a Web search engine and our combined approach for identification of false friends. Section 4 describes the experiments we performed, the resources we used and the results we obtained. It presents a comparison of our different methods for extracting false friends, comparison with previously known algorithms and discussion of the results. Section 5 and Section 6 provide conclusion and discussion of possible future work.

2. Previous work. Previous work on identification of false friends from text corpora could be split in 3 areas: methods for measuring orthographic and phonetic similarity, statistical methods for identification of cognates and false friends from parallel corpora and semantic approaches for distinguishing between cognates and false friends.

Most of the research towards identification of cognates and false friends in the last decade is focused on cognates and primary on orthographic methods for cognate identification which can not distinguish between cognates and false friends. Most studies propose algorithms for extraction of cognates from various sources and using various methods but do not try to distinguish between false friends and other orthographically and semantically non-similar words. Too little attention was given on the problem of distinguishing between cognates and false friends and the task of identification of false friends from parallel corpora.

Traditional orthographic similarity measures like LCSR (*longest common subsequence ratio*) and MEDR (*minimum edit distance ratio*) evolved through the years towards machine learning algorithms for identifying cross-lingual orthographical transformation patterns (like the proposed in [2] and [26]). Recent researchers started using semantic evidence to identification of cognates in addition to the traditional orthographic similarity based algorithms and report improved accuracy ([28] and [26]).

Very little research was conducted on extraction of false friends from parallel corpora. Only few authors (like [29]) proposed such algorithms while many research was conducted on word to word alignment (like [37]) and extraction of bilingual lexicons which can be used for extraction of cognates (like [10] and [24]).

Our approach is a bit different than the outstanding previous research. To extract false friends from parallel corpora we combine statistical techniques observing words occurrences and co-occurrences in a parallel text and techniques for measuring semantic similarity using the Web as a corpus.

2.1. Orthographic and Phonetic Similarity. The first methods proposed for identification of cognates were based on measuring orthographic similarity. For languages sharing the same alphabet classical approaches include measuring Levenshtein minimum edit distance (MED) [22], the longest common subsequence ratio (LCSR) [25] and different variants of Dice's coefficient measuring shared character bigrams [5].

Minimum Edit Distance Ratio (MEDR). The *Levenshtein distance* or *minimum edit distance* (MED) is the minimum number of edit/replace/delete operations of a single character required to transform one string into another [22]. For example transforming the Bulgarian *пърviят* [pərviyat] (*first*) to the Russian *первый* [pervəi:] requires at minimum 4 such operations (replace *ъ* → *е*, replace *и* → *ы*, replace *я* → *й*, and delete *т*).

To measure orthographic similarity the MED is divided on the length of the longer word and is subtracted from 1. This normalization of the MED is called *minimum edit distance ratio* (MEDR):

$$MEDR(s_1, s_2) = 1 - \frac{MED(s_1, s_2)}{\max(|s_1|, |s_2|)}$$

In our example MEDR('*пърviят*', '*первый*') = 1−4/7 = 3/7. The MEDR is a value between 0 and 1 and expresses the orthographic similarity between given two words. Most similar words have high MEDR, near to 1, while non-similar words has low MEDR, near to 0.

Longest Common Subsequence Ratio. The *longest common subsequence ratio* (LCSR) [25] is another example of classical orthographic similarity measure. For given two words LCSR is calculated as the ratio of the length of their longest (not necessarily contiguous) common subsequence (LCS) and the length of the longer word:

$$LCSR(s_1, s_2) = \frac{|LCS(s_1, s_2)|}{\max(|s_1|, |s_2|)}$$

In our example LCS('*пэрвият*', '*первый*') = 3 (the longest common subsequence is '*прв*') and thus LCSR('*пэрвият*', '*первый*') = 3/7.

Shared Bigrams, N-Grams and Dice's Coefficient. Another approach to measuring orthographic similarity between given two words s_1 and s_2 is to calculate the Dice's coefficient for their bigrams [1]:

$$DICE(s_1, s_2) = \frac{2 \times |bigrams(s_1) \cap bigrams(s_2)|}{|bigrams(s_1)| + |bigrams(s_2)|}$$

In the above formula $bigrams(x)$ is a multi-set consisting of all sequences of 2 consecutive characters (bigrams) in the word x. For our example DICE('*пэрвият*', '*первый*') = 2/11.

The idea is further exploited by [5] who introduce few modifications by extending and weighting the Dice's coefficient based formula.

Later, in their study of the identification of confusable drug names [16] develop the idea of using bi-grams further and introduce *tri-grams*, *n-grams*, and a *generalized n-gram measure*. They show that their n-grams based measures BI-SIM and TRI-SIM outperform the traditional orthographics measures like LCSR and MEDR on the test set of the United States Pharmacopeial Convention.

Phonetic Similarity. Phonetic similarity measures how much two words sound alike. Unlike orthographic similarity it measures similarity between the sounds comprising the words, not the letters.

Russel [34] patented a technique to measure phonetic similarity between person names (later called SOUNDEX) based on grouping letters that sound alike and replacing each letter except the first by a code assigned for its group. The algorithm applies also a set of few additional transformation rules and assigns a letter followed by 3 digits to a person name. Persons with similar names were expected to have the same code. For example *Robert* and *Rupert* have the same code *R*163.

Later, many researchers extend and improve the SOUNDEX algorithm. For example [39] combine the letter-grouping idea of SOUNDEX with the minimum edit distance (MED) measure and describe algorithm called EDITEX which assigns smaller weight for replacing letters belonging to the same group.

In our modified minimum edit distance ratio algorithm (MMEDR) [30] we also assign smaller weights to the transformations that replace phonetically similar letters.

Guy [11] describes an algorithm for identification of cognates in bilingual wordlists based on the recurrent sound correspondences. It estimates the probabilities of phoneme correspondences by using a variant of chi-square statistic on a contingency table, which holds how often given two phonemes co-occur in words of the same meaning. The algorithm used only simple on-to-one phoneme correspondences.

The ALINE algorithm ([17]) is capable to identify phonetic similarity between a pair of phonetically transcribed words. It assigns a similarity score to pairs of transcribed words by decomposing phonemes into elementary phonetic features, such as place of articulation, manner of articulation, voice, etc. Features are assigned a weight based on their relative importance. Feature values are encoded as numbers between 0 and 1. The similarity score is then computed by a dynamic-programming algorithm that finds the optimal sequence of operations insert/delete, substitute, and expand/compress.

Following these ideas of Kondrak [19] also used sound correspondence to identify cognates between languages. His algorithm was initially designed for extracting non-compositional compounds from bitexts but it is also able to find complex sound correspondences in bilingual wordlists, not just simple on-to-one phoneme correspondences.

Phonetic similarity can be measured on the basis of the phonetic transcription of the words: first the words are transcribed as a sequence of sounds represented by characters and then the orthographic similarity between these sequences is measured. Transcription allows measuring phonetic similarity between languages using different alphabets. For example in our modified minimum edit distance ratio algorithm (MMEDR) [30] we perform phonetic transcription to replace Russian letters with their Bulgarian equivalents.

Kondrak and Dorr [16] combine several phonetic and orthographic approaches in their study of the identification of confusable drug names and report high accuracy. They conclude that a simple average of several orthographic similarity measures outperforms all individual measures on the task of the identification of confusable drug names.

Manual Transformation Rules. Rather than applying directly some string similarity measure like MEDR or LCSR some studies first apply a set of transformation rules that reflect some typical cross-lingual transformation patterns observed for given pair of languages. This is absolutely necessary when the languages do not use exactly the same alphabet which requires some letters from the first language to be replaced with letters from the second. This idea can be further developed to replace not just single letters but also syllables, endings and prefixes.

For example in [30] we apply a set of manually constructed transformation rules for replacing Bulgarian with Russian endings, replace double consonants with single and replace Russian-specific letters with their Bulgarian equivalents. After that we use a modification of MEDR algorithm that assigns weights for the replace operations reflecting some regular phonetic changes between Bulgarian and Russian.

Manually constructed transformation rules between English and German words (like replacing the letters *k* and *z* by *c* and changing the ending *-tät* by *-ty*) are exploited also by [15] for expanding a list of cognates.

Learning Transformation Rules. The idea of learning automatically cross-lingual transformation rules that reflect the regular phonetic changes between a pair of languages has been exploited by number of researchers. Such techniques follow naturally the idea of using manually constructed transformation rules.

Tiedemann [36] used various measures to learn the regular spelling transformations between English and Swedish from a set of known cognate pairs. His best performing string similarity measure algorithm NMmap uses LCSR algorithm to identify the non-matching parts of two strings and statistically assigns weights corresponding to the probability for transforming between them.

The algorithm proposed by Mulloni and Pekar [27] extracts automatically from a list of known cognates a set of rules that capture regularities in the orthographic transformations between given two languages. These transformations are substitutions of a sequence of letters from the first language with a sequence of letters in the second language identified through the minimum edit distance algorithm. Special characters are added at the word boundaries to allow capturing of rules that transform the start, the middle and the end of the words. For each rule chi-square statistics is calculated and most regular rules are truncated and used while the others are ignored. Finally the transformation rules are applied as a preprocessing step and after that the normalized minimum edit distance is calculated as a similarity measure.

Mitkov et al. [26] use very similar methodology. They collect and score the transformation rules the same way like Mulloni and Pekar [27] but do not account word boundaries as special case. Once the rules are collected and scored by chi-square statistics they apply the rules on candidate pair of words and use LCSR to calculate their similarity.

All of the above techniques use positive examples of cognate pairs to learn regular transformation rules. Unlike them Bergsma and Kondrak [2] use positive and negative examples of cognate pairs to learn positive or negative weights on substring pairings in order to better identify related substring transformations. Starting from minimum edit distance they obtain an alignment of the letters in the given strings and extract corresponding substrings consistent with the alignment. Finally a support vector machine (SVM) is trained by using sets of positive and negative cognate examples and the SVM is used to discriminatively classify given two words as cognates or not.

2.2. Statistical Approach for False Friends Identification. There is no much research concerning extracting false friends directly from text corpora. Most methods (like [26] and [32]) first extract cognates and false friends candidates using some measure of orthographical or phonetic similarity and later try to distinguish between true cognates and false friends.

Fung [10] proposes methods for creating bilingual lexicons from parallel corpora and comparable corpora. His method for extracting semantically related words from sentence level aligned parallel corpus works as follows: for each word pair two binary occurrence vectors are constructed. The first vector maps the occurrences of the first word in the sentences at the left side of the parallel text. The second vector maps the occurrences of the second word in the sentences at the right side of the parallel text. Finally the correlation between these vectors is calculated and used as measure for semantic relatedness.

Brew and McKelvie [5] use sentence alignment to extract cognates and false friends directly from parallel bilingual corpora. The semantic relatedness is identified by statistical method based on collocation analysis in the aligned sentences. The orthographic similarity is measured by various string similarity algorithms. As a result the extracted candidate pairs are classified as cognates, translations, false friends, or unrelated. Their experiments are limited to verbs in English and French but their approach is capable to be applied for other languages as well.

Nakov and Pacovski [29] extract false friends directly from a parallel corpus. Their idea follows the intuition that false friends are unlikely to co-occur

in paragraphs that are translations of each other, while cognates tend to do so. Therefore, good candidates for false friends are words that are frequent in one or both of the languages, but do not co-occur in the corresponding paragraphs or do so rarely, by chance. Based on this idea the authors collect from Bulgarian-Macedonian parallel corpus statistical information about word occurrences and co-occurrences in the corresponding paragraphs and propose several formulas for scoring the likeliness of a pair of identical words to be false friends. Their best performing formula is:

$$F_6(w) = \frac{Par_{BG\&MK}(w) + 1}{\max\left(\frac{1 + Par_{BG}(w)}{1 + Par_{MK}(w)}, \frac{1 + Par_{MK}(w)}{1 + Par_{BG}(w)}\right)}$$

where $Par_{BG}(w)$, $Par_{MK}(w)$ and $Par_{BG\&MK}(w)$ are respectively the number of parallel paragraphs whose Bulgarian side contains the word w, whose Macedonian side contains w, and whose both sides contain w. In their experiments the above formula achieved 85% precision at the top 20 results, and a mean-average precision of 0.562.

2.3. Semantic Approach for False Friends Identification. Methods for measuring semantic similarity are constantly being researched in the last decade. Most of them are based on the *distributional hypothesis* [12] which states that words that occur in similar contexts tend to be similar. A number of methods based on extracting word contexts from various sources have been proposed and studied. Some of them take a window of certain size around the target word ([31]) while others limit the context to words appearing in a certain syntactic relation to the target word such as direct objects of a verb ([23], [28]). A number of methods for comparing word contexts like calculating Dice coefficient ([28]), measuring cosine between vectors ([31]) and many others ([8]) have also been evaluated.

Algorithms extracting semantic similarity based on the distributional hypothesis are proposed by Lin [23] and Curran [8]. In these papers, the contexts are defined based on predefined grammatical relations that are retrieved from a language corpus.

Kondrak [18] propose an algorithm for identifying cognates by combining phonetic and semantic similarity. His system called COGIT has a phonetic module that identifies candidate cognate pairs and a semantic module which judges between cognates and non-cognates. The semantic module uses cross-lingual glossary as bridge between languages and WordNet [9] as source of semantic relatedness between words. Various lexical relations from WordNet like synonymy and hyponymy are exploited.

Kondrak [20] extended his algorithm for measuring semantic similarity based on WordNet and used eight semantic similarity levels as binary features: gloss identity, keyword identity, gloss synonymy, keyword synonymy, gloss hypernymy, keyword hypernymy, gloss meronymy and keyword meronymy. These features are combined with a feature based on phonetic similarity and naive Bayes classifier is used to distinguish between cognates and non-cognates.

Mitkov et al. [26] proposed few methods for measuring semantic similarity between orthographically similar pairs of words used to distinguish between cognates and false friends on the basis of similarity threshold estimated on a training data set. Their first method uses comparable corpora and relies on the distributional similarity. For given pair of words a set of N most similar words are collected using *skew divergence* [21] as similarity function. The similarity between the words is calculated as Dice coefficient between the obtained sets. A bilingual glossary is used to check if two words can be translations of each other. Their second method extracts co-occurrence statistics for each word of interest from the respective monolingual corpus using a dependency parser. Thus verbs are used as distributional features of the nouns. Semantic vectors are created for the two sets of verbs (using *skew divergence* again) and similarity between them is measured by Dice coefficient and using a bilingual glossary. The first method requires a glossary of equivalent nouns while the second requires a glossary of equivalent verbs. In the same study the first method is further extended to use taxonomy data from EuroWordNet (when available). The proposed methods are shown to have different performance on different language pairs and none of them was superior to the others.

The idea of using the Web as a corpus has been exploited by many scientists working on different problems (see [14] for an overview). Some of them use Web search engines for finding how many times a word or phrase is met on the Web and extracting pointwise mutual information ([13]), whereas others directly retrieve context from the text snippets returned by the Web search engines ([31]).

The idea of retrieving information from the text snippets returned by Web search engines is used in [6]. The model they introduce is based on the idea that if two words X and Y are semantically bound, then searching for X should cause Y to appear often in the results, and vice versa: searching for Y should cause X to appear often in the results. As it is later discovered by Bollegala et al. [4], this produces incorrect zero semantic similarity for most of the processed pairs.

Bollegala et al. [4] combine retrieval of information about the number of occurrences of two words (both together and individually) from a Web search

engine, with retrieval of information from the text snippets returned by querying the search engine. They automatically discover lexico-syntactic templates for semantically related and unrelated words using WordNet, and train a support vector machine (SVM) classifier. The learned templates are used for extracting information from the text fragments returned by the search engine and finally, the results are combined.

3. Our Method. We propose a method for extracting pairs of false friends from parallel corpus that combines statistical and semantic evidence for distinguishing between cognates and false friends. We execute two major steps: finding a list of candidate pairs of words and identification of false friends in the list.

3.1. Finding Candidate Cognates/False Friends. The first step we perform aims to find all pairs of words that are perceived as similar and could be cognates or false friends. Given the two texts in Bulgarian and Russian we extract all words from them and for each pair of Bulgarian and Russian word (w_{bg}, w_{ru}) we measure the orthographic similarity and take the pair if the similarity is above given threshold. Because Bulgarian and Russian are highly inflectional languages, we consider all word forms of the same lemma as different words. We don't account part of speech, gender, singular/plural, definite article and case which are expressed as inflections in Bulgarian and Russian.

To measure the orthographic similarity between given pair of Bulgarian and Russian words we use a *modified minimum edit distance ratio* (MMEDR) algorithm described in details in [30]. The MMEDR algorithm first applies a set of manually constructed orthographic transformation rules that replace specific Bulgarian patterns with specific Russian patterns. Later it assigns manually estimated weights to the edit/delete/insert/replace operations and calculates the minimum edit distance between the words. The obtained result is further normalized by dividing to the length of the longer word. Finally the obtained value (which is between 0 and 1) is subtracted from 1 and is used as measure for the orthographic similarity between the words. It has higher value for more similar words and lower – for less similar ones. Although this approach is orthographic, it incorporates also phonetic characteristics because it applies transformation rules and assigns transformation weights motivated by regular phonetic changes between Bulgarian and Russian.

We acknowledge that the MMEDR algorithm can be further improved to automatically learn transformation rules following [27], [2] and [26] but this is out

of scope of the present study. Instead we focus on distinguishing between cognates and false friends which is quite more challengeable task.

3.2. Distinguishing between Cognates and False Friends. The second step we perform aims to distinguish between cognates and false friends. We are particularly interested to identify all false friends in a list of candidate pairs of words. We don't distinguish between true cognates and partial cognates and are only interested in extracting false friends.

3.3. Statistical Approach Our statistical approach for identification of false friends is based on the observations of words occurrences and co-occurrences in the parallel sentences of the corpora we analyze. We follow the basic intuition that in a parallel text cognates tend to co-occur in the corresponding sentences while this is not true for the false friends [29]. To formalize this idea we use the following notations:

- $S_{bg}(w_{bg})$ – the number of Bulgarian sentences in the parallel text containing the word w_{bg}.
- $S_{ru}(w_{ru})$ – the number of Russian sentences in the parallel text containing the word w_{ru}.
- $S_{bg\&ru}(w_{bg}, w_{ru})$ – the number of corresponding sentences in the parallel text containing the word w_{bg} in the Bulgarian sentence and w_{ru} in the Russian sentence.

Following [29] we start by using an adoption of their best performing formula (F_6) to calculate statistically the similarity between a pair of words (w_{bg}, w_{ru}):

$$F_6(w_{bg}, w_{ru}) = \frac{S_{bg\&ru}(w_{bg}, w_{ru}) + 1}{\max\left(\frac{1 + S_{bg}(w_{bg})}{1 + S_{ru}(w_{ru})}, \frac{1 + S_{ru}(w_{ru})}{1 + S_{bg}(w_{bg})}\right)}$$

New Formulas for Statistical Similarity Calculation. Obviously $S_{bg}(w_{bg}) \geq S_{bg\&ru}(w_{bg}, w_{ru})$ and $S_{ru}(w_{ru}) \geq S_{bg\&ru}(w_{bg}, w_{ru})$. Having a high number of co-occurrences $S_{bg\&ru}(w_{bg}, w_{ru})$ should increase the probability that the words $w_{bg}$and w_{ru} are cognates. In the same time having big difference between $S_{bg}(w_{bg})$ and $S_{bg}(w_{bg})$ increases the probability that the words $w_{bg}$and w_{ru} are false friends. Based on these observations we propose two additional formulas (F_1 and F_2):

$$F_1(w_{bg}, w_{ru}) = \frac{(S_{bg\&ru}(w_{bg},w_{ru}) + 1)^2}{(S_{bg}(w_{bg}) + 1)(S_{ru}(w_{ru}) + 1)}$$

$$F_2(w_{bg}, w_{ru}) = \frac{(S_{bg\&ru}(w_{bg}, w_{ru}) + 1)^2}{(S_{bg}(w_{bg}) - S_{bg\&ru}(w_{bg}, w_{ru}) + 1)(S_{ru}(w_{ru}) - S_{bg\&ru}(w_{bg}, w_{ru}) + 1)}$$

Lemmatization. To further improve the accuracy of the statistical method for identification of false friends from a parallel corpus we perform lemmatization. Because Bulgarian and Russian are highly inflectional languages a single word typically has a number of word forms. When calculating $S_{bg}(w_{bg})$, $S_{ru}(w_{ru})$ and $S_{bg\&ru}(w_{bg}, w_{ru})$ we want to consider the same all different forms of given word. We achieve this by applying lemmatization: replace each word with its lemma before counting the occurrences and co-occurrences of the Bulgarian and Russian words. We use large lexicons of lemmas for Bulgarian and Russian. When a word has several lemmas in the lexicon we take into account all of them.

3.4. Semantic Approach. Our semantic approach for distinguishing between false friends and cognates is based on the algorithm described in [31]. The basic intuition used is that if two words are cognates, then most of the words in their respective local contexts should be translations of each other. The idea is formalized using the Web as a corpus, a glossary of known word translations used as cross-lingual "bridges", and the vector space model.

We extract the local context of given word from the text snippets returned by searching in Google. We use as a context all words in a window of size 3 around the target word. We calculate the similarity between given Bulgarian and Russian word by using a glossary of known translation pairs of words. For the Bulgarian word we create a vector of occurrences of all Bulgarian glossary words in the context of the Bulgarian word. For the Russian word we create a vector of occurrences of the corresponding translations of all Bulgarian glossary words into Russian. Finally we calculate cosine between these vectors.

Contextual Web Similarity. We measure the semantic similarity between a Bulgarian word w_{bg} and a Russian word w_{ru} by constructing corresponding contextual semantic vectors V_{bg} and V_{ru} and comparing them through the glossary G of translation pairs.

The process of building V_{bg}, starts with a query in Google limited to Bulgarian pages for the target word w_{bg}. We collect the resulting page titles and text snippets (up to 1 000), and we remove all stop words (prepositions, pronouns, conjunctions, interjections and some adverbs) and words shorter than 3 letters. We replace all uppercase letters with their corresponding lowercase letters.

We then identify all occurrences of w_{bg} or one of its word forms (using the lexicon of lemmas) in the page titles and text snippets returned by Google and we extract 3 words on either side of each occurrence. Finally, for each collected word, we calculate the number of times it has been extracted, thus producing a contextual semantic frequency vector V_{bg}.

For example let's assume we want to calculate the semantic context vector V_{bg} for the Bulgarian word *картина* [kartina] (painting). We perform search in Google for *картина* specifying to search Bulgarian pages only and collect all returned page titles and text snippets (see Table 1).

Table 1. Results of searching the Bulgarian word картина in Google

Нощна стража (**картина**) — Уикипедия
В момента **картината** е изложена в музея Рейксмузеум в Амстердам. Истинското име на **картината** е "Ротата на капитан Банинг Кок". Тъй като престояла дълги ...
Картина с известни личност | спанак.орг
Огромна **картина**, на която са изобразени много известни личности – Айнщайн, Чърчил, Линкълн, Фидел Кастро, Че Гевара. От новата вълна можете да намерите ...
Намерена е най-древната картина в света – MystiColors Forum
В будисткия комплекс Бамиян (Bamiyan) в Афганистан група японски археолози намериха най-древната в света **картина**, нарисувана с маслени бои. ...
...

We remove all stop words and words with length less than 3 and replace all uppercase letters with their corresponding lowercase letters. We replace all words with their corresponding lemmas (apply lemmatization). Finally we extract all words in a window of size 3 around each occurrence of *картина*. As a result we obtain the semantic context vector V_{bg} (Table 2).

Similarly we repeat the procedure for w_{ru} to obtain a Russian contextual semantic frequency vector V_{ru}. Once we have the contextual vectors V_{bg} and V_{ru} we need to measure similarity between them. For the Bulgarian word w_{bg} we create a vector G_{bg} containing the number of occurrences in V_{bg} of each Bulgarian glossary word. For the Russian word w_{ru} we create a vector G_{ru} containing the total number of occurrences in V_{ru} of the translations of each Bulgarian glossary word into Russian. The vectors G_{bg} and G_{ru} have the same size – the number of Bulgarian words in G. For each Bulgarian word w from G we have a corresponding entry in G_{bg} and in G_{ru} that show how many occurrences of w exist in V_{bg} and respectively how many occurrences of translations of w into Russian exist in V_{ru}.

Table 2. The semantic context vector V_{bg} containing the context words and their corresponding number of occurrences extracted for the Bulgarian word картина from Google

word	occurences
картина	461
купувам	386
скъп	345
известен	205
галерия	183
голям	176
изкуство	188
художник	98
рисувам	91
фотоапарат	2
...	...

Table 3. Vectors G_{bg} and G_{ru} and their corresponding words from G (with abridgements)

Bulgarian word from G	G_{bg}	G_{ru}
абитуриент (school leaver)	0	0
абонамент (subscription)	2	0
абонат (subscriber)	0	0
...	...	...
галерия (gallery)	94	143
голям (big)	56	176
известен (famous)	84	205
изкуство (art)	167	188
картина (painting)	262	461
купувам (buy)	72	96
рисувам (paint)	202	171
скъп (expensive)	133	45
фотоапарат (camera)	0	2
художник (painter)	122	398
...	...	...

For example let's assume w_{bg} is *картина* (painting) and w_{ru} is *художник* (painter). We obtain vectors G_{bg} and G_{ru} as follows (see Table 3).

Finally we calculate the cosine between the vectors G_{bg} and G_{ru} and thus we obtain a number between 0 and 1 corresponding to the semantic similarity between w_{bg} and w_{ru} (higher value means more similar words) calculated by using the Web as a large monolingual corpus (for Bulgarian and for Russian separately).

3.5. Combined Approach. Both statistical and semantic approach can distinguish between cognates and false friends with satisfying accuracy but both of them have weak sides that can be improved.

The statistical approach works well when we have rich statistics for given two words but when the words appear in the text too little number of times, the accuracy of the statistics is not good. For example if a pair of words w_{bg} and w_{ru} appear only 1-2 times in the text and appear once in corresponding sentences, the algorithm will be unsure to decide whether these words are cognates or false friends. Occurrences of words and co-occurrences of words in corresponding sentences could happen by chance if the words appear in the text only 1–2 times. In the opposite case when we have words appearing 50–60 times in the text, the statistics for their occurrences and co-occurrences is rich and the algorithm will distinguish accurately between false friends and cognates (true friends).

The semantic approach works differently and it gathers information about the pair of words only from the Web. Its accuracy is generally good for words which are entirely different but sometimes it assigns very low values for highly related words. There are different reasons for inaccuracy of the semantic approach. The main problem comes because it relies on the Google search engine which returns only the first 1 000 matches when searching for given word and it rates higher news sites, e-commerce sites and blogs, which distorts the extracted local contexts. Some words related to geographical and cultural particularities have different contexts on the Web for Bulgarian and Russian while generally are highly related. Good examples are person names and names of goods used in e-commerce (due to different popular brands in different countries).

Combining the statistical and semantic approaches is natural because the statistical approach returns similarity values between 0 and 1 for words that do not have rich statistics collected and the confidence in such cases is not good. The statistical approach gives high values (above 1) for words that are highly related and this conclusion is based on rich statistics. In the same time the semantic approach gives high values (near to 1) for highly related words and low values for unrelated words (near to 0). Consequently combining the two approaches by simple summing of the values returned by each of them seems natural. Our experiments confirm that such way of combining the methods is valuable and increases the accuracy of the results.

We also tried combining the statistical and semantic approach by weighting their score and we found that weighting does not yield significant improvement of the results.

4. Experiments and Evaluation. We performed multiple experiments to measure the performance of the described algorithms and combinations of them. We used a sentence-level aligned parallel corpus – a portion of the Russian novel "Lord of the World" by Alexander Beliaev and its Bulgarian translation consisting of 759 parallel sentences.

As a first step we extracted the pairs of words that are perceived as similar and should be recognized as cognates or false friends. For the extraction we used the MMEDR algorithm (described in details in Section 3.1) with threshold of 0.90. As a result we got 612 candidate pairs of words which were judged by a linguist as false friends/not false friends (which include partial cognates and true cognates). False friends were 35 of them (5.72%), partial cognates were 67 (10.95%) and true cognates were 510 (83.33%).

As a second step we applied several algorithms to distinguish between false friends and cognates. All of them produced a list consisting of all the pairs identified as candidates at the previous step ordered by their similarity calculated by the respective algorithm. The false friends were expected to be in the beginning of the list (having similarity near to 0), followed by the cognates. The algorithms do not distinguish between partial cognates and true cognates. Following [2] and [31] the evaluation were performed by using the well-known in information retrieval measure 11-pt average precision which averages the precision at 11 points corresponding to recall of respectively 0%, 10%, 20%, ..., 100%.

We experimented with the statistical approach for identification of false friends (described in details in Section 3.3) with and without lemmatization and using different formulas to compute the similarity from the occurrences and co-occurrences of the words in the parallel text. We also experimented with the semantic algorithm for identification of false friends (described in details in Section 3.4). Finally we combined the statistical and semantic approaches in a new improved algorithm and compared it with the others. All experiments and algorithm parameters are described below (in Section 4.2).

4.1. Resources. For the purpose of the experiments and implementation of the algorithms we used the following resources: parallel corpus, lemmatization lexicons and bilingual glossary.

Sentence-Level Aligned Parallel Corpus. We used the first 7 chapters of the Russian novel "Lord of the World" by Alexander Beliaev and its Bulgarian translation consisting of 759 aligned parallel sentences from which we extracted 612 pairs of words candidates for classification as cognates or false friends.

Lemmatization Lexicons. We used two large monolingual morphological lexicons for lemmatization for Bulgarian and Russian.

The Bulgarian morphological lexicon [33] is created at the Linguistic Modeling Department of the Institute for Parallel Processing in the Bulgarian Academy of Sciences (BAS) and contains about 1 000 000 wordforms and 70 000 lemmata. Each lexicon entry consists of a wordform, a corresponding lemma, followed by morphological and grammatical information. There can be multiple entries for the same wordform, in case of multiple homographs.

The Russian morphological lexicon [33] is also created at the Linguistic Modeling Department of the Institute for Parallel Processing in the Bulgarian Academy of Sciences (BAS). It is in the same format like the Bulgarian and contains about 1 500 000 wordforms and 100 000 lemmata. Its core content is based on the grammatical dictionary of [38].

Bilingual Glossary. We used a large Bulgarian-Russian electronic glossary consisting of 59 582 pairs of words which are translations of each other. The glossary was adopted by scanning, parsing and processing the Bulgarian-Russian dictionary of [3] and the Russian-Bulgarian dictionary of [7]. We use the word-word translations from these dictionaries ignoring the phrase-word and phrase-phrase translations. Most of the words have multiple translations so we have a set of Russian translation words for each Bulgarian word and vice versa. This is taken into account during the comparison of the Bulgarian and Russian contextual semantic vectors as described in Section 3.4.

Searches in Google. During our experiments we performed searches in Google for 557 Bulgarian and 550 Russian wordforms and collected as many as possible (up to 1000) page titles and text snippets from the search results. We used this text information to extract the local contexts of these words and build their contextual semantic vectors as described in Section 3.4.

4.2.Experiments. This section describes the experiments performed with the statistical, semantic and combined algorithms for identification of false friends.

Baseline. As baseline we took the following algorithm:

- **ASC** – words pairs sorted in ascending order (first by the Bulgarian word and second by the Russian word). It behaves nearly like a random function.

Statistical Algorithms. We performed the following experiments based on the statistical approach for identifying false friends in a parallel text:

- **PAR** – the original algorithm of Nakov and Pacovski [29] (without lemmatization) with their formula F_6.
- **PAR+L** – the algorithm PAR, modified to use lemmatization.
- **F1** – the algorithm PAR applied with the formula F_1.
- **F1+L** – the algorithm PAR applied with the formula F_1.
- **F2** – the algorithm PAR applied with the formula F_2.
- **F2+L** – the algorithm PAR, with the formula F_2 and with lemmatization.

Semantic Algorithms. We performed the following experiments exploiting the semantic approach for identification of false friends:

- **SIM** – the algorithm for extraction of semantic similarity from the Web with lemmatization. It does not use the statistical information about occurrences and co-occurrences of the words in the parallel corpus.

Combined Algorithms. We also tried different ways of combining the semantic and statistical algorithms resulting in the following experiments:

- **SIM+F1+L** – the algorithm SIM combined with the algorithm F1+L by summing the values of SIM and F1+L.
- **SIM+F2+L** – the algorithm SIM combined with the algorithm F2+L by summing the values of SIM and F2+L.
- **1.5∗SIM+F1+L** – the algorithm SIM combined with the algorithm F1+L by weighted summing the values of 1.5∗SIM and F1+L.
- **1.5∗SIM+F2+L** – the algorithm SIM combined with the algorithm F2+L by weighted summing the values of 1.5∗SIM and F2+L.
- **SIM+1.5∗(F1+L)** – the algorithm SIM combined with the algorithm F1+L by weighted summing the values of SIM and 1.5∗(F1+L).
- **SIM+1.5∗(F2+L)** – the algorithm SIM combined with the algorithm F2+L by weighted summing the values of SIM and 1.5∗(F2+L).

Table 4. Comparison of the evaluated algorithms for identification of false friends

Algorithm	11-pt average precision
ASC	4,17%
F2	38,60%
F1	39,50%
PAR	43,81%
PAR+L	53,20%
SIM	63,68%
F1+L	63,98%
F2+L	66,82%
SIM+1.5∗(F2+L)	74,34%
1.5∗SIM+F1+L	75,07%
SIM+1.5∗(F1+L)	75,46%
SIM+F2+L	76,15%
SIM+F1+L	77,50%
1.5∗SIM+F2+L	77,64%

4.3. Results. The Table 4 summarizes the results obtained by the evaluated algorithms (ordered from the worst to the best).

4.4. Discussion. The results show good level of accuracy of the best performing algorithms far away from the baseline. It is obvious that for Bulgarian and Russian which are highly inflectional languages, applying lemmatization is a must. When combined with lemmatization our new formulas F_1 and F_2 perform significantly better than the original formula F_6 (the PAR algorithm) taken from [29]. All combined methods perform better than the statistical and the semantic approach individually. Weighting the statistical score and semantic score in the combined algorithms almost does not yield improvement.

Generally the proposed algorithms are applicable for other language pairs, different than Bulgarian and Russian. The required resources are parallel text, bilingual glossary and ability to perform search queries to Google (which needs to support the target languages). In a previous study ([31]) we have shown that significantly smaller glossary (about 4500 words) can be used and this has almost zero impact over the accuracy. Queries to Google can be done only once and the results can be stored as a cache to allow reuse. The algorithms need at most about 2∗10 queries per word pair that is classified as cognate or false friend so it is not expensive. Lemmatization lexicons are be required only if we process highly inflectional languages.

Our algorithms does not distinguish between different parts of speech and

identifies as candidate cognates all words that are similar enough by the MMED algorithm, despite of the fact that different part of speech in most cases are false friends even when are strongly related semantically. For example a verb and adjective could not be cognates or partial cognates. Orthographically identical and similar prepositions in most cases are partial cognates because they always have multiple translations and some of them are mutual translations but some of them differ. Orthographically identical pronouns between Bulgarian and Russian in most cases have different meaning and are behaving as false friends. Most of these regularities are identified correctly by our combined algorithms.

5. Conclusion. We proposed an algorithm for extracting false friends from a sentence level aligned parallel text that combines statistical and semantic evidence for distinguishing between cognates and false friends. Our algorithm improves significantly the existing pure statistical approaches and shows that false friends can be efficiently extracted from parallel texts without human supervision. The proposed use of the Web as a corpus to distinguish between cognates and false friends is a promising novel approach that can be further improved and combined with other semantic methods.

6. Future Work. Generally, we have significant improvement over the original statistical algorithm of [29] but our results are still not perfect. We want to try different improvements of the statistical, semantic and combined algorithms.

We would like to improve the formulas for measuring semantic similarity based on the occurrences and co-occurrences of the words in a parallel text. The approach of assigning a mapping vectors to the occurrences of each word in the sentences and calculating cosine between the vectors of other similarity measure (as in [10]) is also not evaluated.

Later we want to try using non-parallel corpus and extracting distributional similarity as it was shown in [26].

We want to improve the semantic algorithm for measuring semantic similarity through the Web by using certain syntactic relations between the words when extracting the local context. Our current approach takes all words in a window of few words around the target word to build its local context vector but this could be potentially improved following [26] and [4] and using only specific syntactic relation to the target word such as direct objects of a verb after applying dependency parsing.

We would like to try adding taxonomic evidence for identification of false friends by using various resources like WordNet, EuroWordNet and other taxonomies. This has never been done for Bulgarian and Russian so it is a challengeable task.

Finally we would like to implement the algorithm for different language pairs and to compare the results with other algorithms for extraction of false friends, such like [26].

REFERENCES

[1] Adamson G., J. Boreham. The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles. *Information Storage and Retrieval*, **10** (1974), 253–260.

[2] Bergsma S., G. Kondrak. Alignment-Based Discriminative String Similarity. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 2007, 656–663.

[3] Bernstein B. Bulgarian-Russian Dictionary. Moscow, Russia, 1986. (Бернщайн С. Б. Болгарско-русский словарь, изд. Русский язык, Москва, 1986).

[4] Bollegala D., Y. Matsuo, M. Ishizuka. Measuring Semantic Similarity between Words Using Web Search Engines. In: Proceedings of the 16th International World Wide Web Conference (WWW2007), Banff, Canada, 2007, 757–766.

[5] Brew C., D. McKelvie. Word-Pair Extraction for Lexicography. In: Proceedings of the 2nd International Conference on New Methods in Language Processing, Ankara, Turkey, 1996, 45–55.

[6] Chen H., M. Lin , Y. Wei. Novel Association Measures Using Web Search with Double Checking. In: Proceedings of the COLING/ACL 2006, Sydney, Australia, 2006, 1009–1016.

[7] Chukalov K. Russian-Bulgarian Dictionary. Moscow, Russia, 1986. (Чукалов С. К. Русско-болгарский словарь, изд. Русский язык, Москва, 1986)

[8] Curran J., M. Moens. Improvements in Automatic Thesaurus Extraction. In: Proceedings of the Workshop on Unsupervised Lexical Acquisition, SIGLEX 2002, Philadelphia, PA, USA, 2002, 59–67.

[9] Fellbaum C. (Ed.) WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, USA, 1998.

[10] Fung P. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. In: Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, Springer, 1998, 1–16.

[11] Guy J. An Algorithm for Identifying Cognates in Bilingual Wordlists and Its Applicability to Machine Translation. *Journal of Quantitative Linguistics*, **1** (1994), No. 1, 35–42.

[12] Harris Z. Distributional Structure. In: Katz J. (editor), The Philosophy of Linguistics, Oxford University Press, New York, NY, USA, 1985, 26–47.

[13] Inkpen D. Near-Synonym Choice in an Intelligent Thesaurus. In: Proceedings of the NAACL-HLT, New York, NY, USA, 2007.

[14] Kilgarriff A., G. Grefenstette. Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics*, **29** (2003), Issue 3, 333–347.

[15] Koehn P., K. Knight. Learning a Translation Lexicon from Monolingual Corpora. In: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, PA, USA, 2002, 9–16.

[16] Kondrak G., B. Dorr. Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In: Proceedings of COLING 2004, Geneva, Switzerland, 2004, 952–958.

[17] Kondrak G. A New Algorithm for the Alignment of Phonetic Sequences. In: Proceedings of NAACL/ANLP 2000: 1st conference of the North American Chapter of the Association for Computational Linguistics and 6th Conference on Applied Natural Language Processing, Seattle, WA, USA, 2000, 288–295.

[18] Kondrak G. Identifying Cognates by Phonetic and Semantic Similarity. In: Proceedings of the 2nd meeting of the North American Chapter of the

Association for Computational Linguistics (NAACL 2001), Pittsburgh, PA, USA, 2001, 288–295.

[19] Kondrak G. Identifying Complex Sound Correspondences in Bilingual Wordlists. In: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003), Mexico City, Mexico, 2003, 432–443.

[20] Kondrak G. Combining Evidence in Cognates Identification. In: Proceedings of the 17th Canadian Conference on Artificial Intelligence, London, 2004, 44–59.

[21] Lee L. Measures of Distributional Similarity. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, USA, 1999, 25–32.

[22] Levenshtein V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Doklady Akademii Nauk SSSR, **163** (1965), No. 4, 845–848.

[23] Lin D. Automatic Retrieval and Clustering of Similar Words. In: Proceedings of COLING-ACL'98, Montreal, Canada, 1998, 768–774.

[24] Mann G., D. Yarowsky. Multipath Translation Lexicon Induction via Bridge Languages. In: Proceedings of NAACL 2001, Pittsburgh, PA, USA, 2001, 151–158.

[25] Melamed D. Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics*, **25** (1999), No. 1, 107–130, ISSN:0891-2017.

[26] Mitkov R., V. Pekar, D. Blagoev, A. Mulloni. Methods for Extracting and Classifying Pairs of Cognates and False Friends. *Machine Translation*, **21** (2007), Issue 1, Springer Netherlands, 29–53.

[27] Mulloni A., V. Pekar Automatic Detection of Orthographic Cues for Cognate Recognition. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06), Genoa, Italy, 2006, 2387–2390.

[28] Mulloni A., V. Pekar, R. Mitkov, D. Blagoev. Semantic Evidence for Automatic Identification of Cognates. In: Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, Borovets, Bulgaria, 2007, 49–54.

[29] Nakov P., V. Pacovski. Acquiring False Friends from Parallel Corpora: Application to South Slavonic Languages. In: Readings in Multilinguality, Selected Papers from Young Researchers in BIS-21++ (Eds M. Slavcheva, G. Angelova, K. Simov) INCOMA Ltd., Shoumen, Bulgaria, 2006, 87–94.

[30] Nakov P., S. Nakov, E. Paskaleva. Improved Word Alignments Using the Web as a Corpus. In: Proceedings of RANLP'2007, Borovets, Bulgaria, 2007, 400–405.

[31] Nakov S., P. Nakov , E. Paskaleva. Cognate or False Friend? Ask the Web! In: Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, Borovets, Bulgaria, 2007, 55–62.

[32] Nakov S. Automatic Acquisition of Synonyms Using the Web as a Corpus. In: Proceedings of the 3rd Annual South-East European Doctoral Student Conference (DSC 2008), Vol. **2**, Thessaloniki, Greece, 2008, 216–229.

[33] Paskaleva E. Compilation and Validation of Morphological Resources. In: Workshop on Balkan Language Resources and Tools (Balkan Conference in Informatics), Thessaloniki, Greece, 2003, 68–74.

[34] Russel R. U.S. Patent 1,261,167, Pittsburgh, PA, USA, 1918.

[35] Simard M., G. Foster, P. Isabelle. Using Cognates to Align Sentences in Bilingual Corpora. In: Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada, 1992, 67–81.

[36] Tiedemann J. Automatic Construction of Weighted String Similarity Measures. In: Proceedings of the SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, USA, 1999, 213–219.

[37] Tiedemann J. Word to Word Alignment Strategies. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, 2004, 212–218.

[38] Zaliznyak A. Grammatical Dictionary of Russian. Moscow, Russia, 1977. (Зализняк А. Граматический словарь русского языка, изд. Русский язык, Москва, 1977).

[39] Zobel J., P. Dart. Phonetic String Matching: Lessons from Information Retrieval. In: Proceedings of the 19th International Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 1996, 166–172.

Svetlin Nakov
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, J. Bourchier Blvd
1164 Sofia, Bulgaria
e-mail: nakov@fmi.uni-sofia.bg

Received March 3, 2009
Final Accepted April 29, 2009