The Role of Homograms in Machine Translation

Lucia Nacinovic Prskalo and Marija Brkic Bakaric

***Abstract*—The Croatian language is a pitch-accent language, in which the tone contour realized in the stressed syllable carries the lexical information. Therefore, in some cases, a different lexical accent gives the word a different meaning. In such cases, the ambiguity of the word in written texts, where accents are not usually marked, can be solved by determining the appropriate accent. There are also cases when various basic and derived forms of words have different meanings, different morphosyntactic descriptions (MSDs), and possibly different accents. When words have the same written forms but different meanings, they are called homograms. In order to resolve the ambiguity of homograms, we created a lexicon of homograms that is comprised of all Croatian nouns of different gender, which have the same written forms (if accents are not marked) but different meanings, MSDs, and possibly different accents. This lexicon consists of 19,366 entries and 3,460 unique homograms. Each entry in the lexicon comprises the homogram (unaccented word), the accented word, the corresponding MSD, and the accented lemma. The obtained lexicon enables us to identify and disambiguate homograms within the corpus efficiently and accurately. We also evaluated and analyzed the performance of machine translation (MT) systems for the Croatian–English language pair with a special emphasis on homogram translation. We confirmed that the disambiguation of homograms can improve the performance of MT systems in avoiding major translation mistakes related to assigning the wrong meaning to homograms.**

***Index Terms*—Disambiguation of homograms, lexicon of homograms, pitch accent language, word sense disambiguation.**

I. Introduction

The aim of this paper is to carry out the task of word sense disambiguation (WSD) in a pitch-accent language using morphosyntactic descriptions (MSDs) and word accents and to assess the performance of machine translation (MT) systems for the Croatian–English language pair to check whether disambiguation of homograms could improve the performance of MT systems in avoiding major translation mistakes related to assigning the wrong meaning to homograms.

Although there are some universal prosody patterns that can be found in most languages, the use of prosodic features is largely specific to different languages. One of the most important prosodic features is tone. According to the way of using tonal variations (pitch), languages can be categorized into tonal and nontonal languages. In nontonal languages, prominence (accent) does not have tonal characteristics, and it is associated only with a certain syllable in a word. In tonal languages, the same sound, pronounced with different tones (pitch), can express different meanings (i.e., lexical or grammatical contrasts) [1]. In most such languages, each syllable in a word can have an independent tone. Tonal languages are highly common in Africa (Bantu, Khoisan, Niger-Congo, and Nilo-Saharan languages) and East Asia (Mandarin Chinese, Vietnamese, and Thai) [2]. Besides stress-accented (nontonal) languages and tonal languages, there is one more category in the typology of accentuated systems—restricted-pitch languages or pitch-accent languages [3]—in which the tone contour realized in pronounced words carries lexical information, but the potentially distinctive tones are mostly restricted to one syllable within a word. Pitch-accent languages can be found in some parts of Europe (Swedish, Norwegian, Lithuanian, Latvian, Slovene, Bosnian, Serbian, and Croatian) and East Asia (Japanese and Korean) [4], [5]. As noted, the accentuated system of the Croatian language belongs to the third category.

The accent is one of the most important prosodic features. Besides having a great role in emphasizing words, accents also have a role in perceiving the boundaries between words in speech. While in the written text the boundaries of the words are clearly seen, the boundaries between the words in speech cannot seem to be perceived so easily, except in those places where speech breaks are present. However, if people cannot recognize the words, they will not be able to understand the message being transmitted. It is therefore clear that we possess the ability to distinguish words in speech, precisely because of the accent and other prosodic features. The accents of words, therefore, play a very important role in communication because they help the listener perceive the word boundaries.

Accents also have a distinctive role in cases where words with the same written form are pronounced differently because of the different accent. Therefore, in some cases, a different lexical accent gives the word a different meaning; for example, in Croatian, the word pas can be *pȁs* (En. dog) and *pâs* (En. waist). In such cases, the ambiguity of the word *pas* in written texts, where accents are not usually marked, can be solved by determining the appropriate accent.

In this paper, we use the Croatian lexicon [6] that contains Croatian words with their appropriate accents and MSDs for disambiguation of the words that have the same written form in text but different MSD and possibly different accent type.

WSD, as one of the tasks of natural language processing (NLP), can be applied in various other NLP tasks to improve their results. Some areas in which WSD can be applied are, for example, MT [7]-[9], information retrieval [10], and text categorization [11]. WSD for MT is a subset of a general

Manuscript received February 7, 2018; revised March 23, 2018. This work has been fully supported by the University of Rijeka under the projects number 13.13.1.3.03 and 16.13.2.2.01.

The authors are with the Department of Informatics, University of Rijeka, Croatia (e-mail: lnacinovic@inf.uniri.hr, mbrkic@inf.uniri.hr).

doi: 10.18178/ijmlc.2018.8.2.669

WSD problem. It is easier in that distinct source-language senses, which share the same translation equivalent, do not need to be differentiated [12]. That phenomenon, known as parallel polysemy, is particularly common among related languages [13]. In phrase-based statistical MT, which is still a very popular approach to MT, extra linguistic information is incorporated with factored translation models [14]. The performance of WSD from the MT perspective is analyzed in three different ways in [7]. Results are evaluated manually by analyzing ambiguous homograms and their suggested translation equivalents and also by analyzing the agreement between the WSD proposed equivalent and those suggested by the systems and by computing the selected MT automatic scores for all translation versions. In manual evaluation, the authors distinguish the following categories: correct, incorrect, and borderline, where borderline refers to semantically correct equivalents which do not have a correct form. The authors in [15] evaluated several options of using verb senses in the source-language sentences as an additional factor for the Moses statistical MT system. The evaluation was performed both manually and automatically. Even though a new approach to MT, that is, neural MT (NMT), can provide a much more flexible mechanism for adding extra linguistic information, the authors in [8] showed that linguistic features are not redundant in NMT. Furthermore, WSD remains a challenging problem even with NMT, especially for rare word senses [9]. The authors in [9] take two approaches to integrate sense embeddings in order to improve WSD in NMT. In their first approach, they passed sense embeddings as additional input to the NMT system. For the second experiment, they extracted lexical chains based on sense embeddings from the document and integrated this information into the NMT model.

WSD did not get much attention for the Croatian language. So far, there are only several papers on WSD in Croatian. Those include [16] where the authors experimented with supervised learning methods in order to determine the position of strong predictors for WSD of Croatian nouns. The authors concluded that words in the immediate vicinity of an observed lexeme (1–5 words left and right) have the highest discriminative power. In [17], the authors investigated the use of active learning for Croatian WSD, and in [18] the authors extended the lexical sample dataset for Croatian WSD which consists of 36 words. The authors also evaluated different WSD models on the created datasets. All the above-mentioned researches did not deal with the ambiguity of homograms, which is the main topic of this research. To the best of our knowledge, this is the first paper that concerns the disambiguation of the homograms in Croatian or any other South Slavic language.

The rest of the paper is organized as follows. In the next section, we review the properties of the Croatian accentuated system. In Section III, we describe the Croatian accent lexicon that we used as a main resource in our research. In Section IV, we explain the ambiguity of homograms, and in Section V we explain the procedure of obtaining the lexicon of homograms and experiments of identifying and disambiguating homograms within a corpus. Manual evaluation of ambiguous homograms and their MT translations is presented in Section VI. Lastly, Section VII concludes the paper and outlines the future work.

II. Croatian Accentuated System

One of the most important prosodic features of the Croatian language is its accentuated system. Pronounced words in Croatian have different accent features that include the intensity of speech, the movement (rising and falling) of the tone in speech, and the duration of a syllable. The vowel in the accented syllable differs from the vowel in the unaccented syllables precisely by those features. According to the definition, accent is a simultaneous realization of the intensity, tone, and duration [19], and prominence of a syllable in a word that has those features represents a word accent. The feature of the intensity is realized as a difference in the power of the sound current, and it is related to the greater or lesser consumption of air from the lungs and pressure that is formed when air travels through the speech organs. The accented syllables are reflected in greater consumption and pressure of air than the unaccented syllables. The moving of the tone (pitch) refers to its falling or rising curve in an accented syllable, and the duration refers to the difference in the length of the vowel in a pronounced syllable. The vowels in the accented syllables are longer than the vowels in the unaccented syllables. According to the mentioned features, there are four types of accents in the Croatian accentuated system: short falling as in the word *kùća* (En. house), short rising as in the word *rúka* (En. hand), long falling as in the word *zlâto* (En. gold), and long rising as in the word *žèna* (En. women) [19].

In some languages, the exact place of the stressed syllable in the word is precisely defined. In French, for example, it is always the last syllable in the word; in Polish and Czech, it is always the first syllable. In some languages, the place of the stressed syllable is not uniquely determined. Such languages are, for example, English and Croatian.

Accent Distribution

In one spoken word in Croatian, only one syllable can be accentuated. As mentioned before, the place of the accent in a word in Croatian is free; it can be on any syllable of a word except on the last one. The accent is most common on the first syllable of the word (in about 66% of the words), then on the second syllable (in about 23% of the words), the third (in about 6.7% of the words), and the fourth (in about 1.6% of the words) [20]. In addition to the fact that the place of the accent is free within a word, it can also be changed within the paradigm, so in the different forms of the same word, all four accents can be found, for example, in the word *lònac* (En. pot), *lònac*, gender: male, number: singular, case: nominative; *lónca*, gender: male, number: singular, case: genitive; *lônče*, gender: male, number: singular, case: vocative; *lònācā*, gender: male, number: plural, case: genitive.

The free distribution of accents in words in Croatian can be limited by the following rules [19]-[21]:

(1) The ascending accents stand only on the first syllable in the word.
(2) In one-syllable words, there are only descending accents.
(3) There is no accent on the last syllable in the word.
(4) The syllables before the accented syllables can be only short, not long.

However, there are categories of words that can be

excluded from the rules:

(1) Some complex words (e.g., poljopr̈ivreda, En. agriculture).
(2) Words of foreign origin (e.g., dirigẽnt, En. bandmaster).
(3) Foreign names (e.g., Voltêr, Montevidẽo).
(4) Abbreviations that are pronounced by naming the initial letters (e.g., /esadê/).
(5) In some words, the ascending accent can stand on the final syllable or in one-syllable words (e.g., bic ȁkl, En. bicycle).

III. Croatian Accent Lexicon

As mentioned in the Introduction, the Croatian language is a pitch-accent language in which the tone contour realized in the stressed syllable carries the lexical information. Therefore, a lexicon with all Croatian words (in their basic and derived forms) and their corresponding accents is an important resource for various NLP areas, such as MT, WSD, and speech synthesis. Such a lexicon was created in [6] by implementing the rules for constructing derived forms of words on the basis of the addition of the appropriate extension and adding the corresponding accent on the syllable of a certain word. The entries in the lexicon are comprised of all basic and derived words written without and with its corresponding accent and MSD or part-of-speech (POS) tag. Altogether, the lexicon is comprised of 72,366 words in their basic form and over 1,000,000 derived word forms.

Croatian is a morphologically very rich language and the rules for constructing derived word forms are rather complex. Those rules have been applied in order to create the Croatian accent lexicon. In her doctoral thesis, Mikelić Preradović [22] developed models and outlined the rules and exceptions for the automatic generation of accented forms of words in all of their forms. In [6], those rules were used and implemented in a programming language in order to create the accent lexicon. The basic forms of the words were obtained from the lexicon "*Veliki rječnik hrvatskoga jezika*" [23], and on their basic form the appropriate paradigmatic sequences and accents were added in order to get all the derived forms of the words with their corresponding accents.

The statement that the Croatian morphology is very rich can easily be supported by the fact that there are, for example, for nouns, two types, three genders, two numbers, seven cases, and an attribute animate which can take values of yes or no. Verbs have two types, seven forms, three persons, two numbers, three genders, and an attribute negative which can take values of yes or no. Other types of words such as adjectives, pronouns, numbers, and adverbs also have many derived forms. There are also some types of words in Croatian that can have more attributes, but they only have one form. Such words are, for example, prepositions, conjunctions, particles, and interjections. The detailed MSDs and their attributes and possible values for the Croatian language can be found in [24].

Accent features of the pronounced words can therefore be studied from the point of view of morphology and word formation, so we can speak about the accent of the singular and plural nouns and the accent of the present verb form, comparative adjective form, and so forth. Such lexical accents are sometimes also called morphological accents.

The rules outlined in [22] refer to the largest groups of words in Croatian: nouns, verbs, and adjectives. When creating the Croatian accent lexicon [6], all other word types were obtained from [23]. Additionally, for some word types that, along with nouns, verbs, and adjectives, can also have different forms, derived forms were added manually in the accent lexicon (e.g., derived forms of most of the pronouns and derived forms of numbers).

Morphosyntactic Description (MSD)

POS tags are associated with the grammatical categories of words in the text. In languages that are not morphologically rich (e.g., English or French) for some NLP tasks, it is sufficient to determine the category of a word (and mark it with the POS tag). Since Croatian is a morphologically very rich language and has numerous word forms, assigning POS tags is not enough to determine the grammatical category of a word. For such languages, in addition to the word type, other MSDs have to be added. When creating the Croatian accent lexicon [6], along with the derivative forms of the words, for each entry in the lexicon, the corresponding MSD was also added. The MSDs followed the MULTEXT-East Standards [25] for the Croatian language. Some attributes of the tags are missing because they could not be determined according to the rules that were taken into consideration when constructing the derived forms of the words. For example, for nouns, all attributes were added except for type, which can take values common or proper and for some nouns attribute animate which can take values of yes and no. Those attributes could not be added because they were not covered by the rules, since changing the values of those two attributes does not change the word form.

IV. Ambiguity of Homograms

In written texts in Croatian, there are no accents marked. Therefore, there can be words and their forms that have the same sequence of graphs in writing but different meaning. Such words are called homograms (which are similar to but not completely the same as homographs in English). If in such words the same sequence of prosody is present, too, that is, they are equally spoken, then they are called homophones, and if they are pronounced differently, then they are called heterophones [26]. Accents have an important distinctive function in heterophonic homograms that have the same graphemic and different prosodemic sequences. So, for example, the Croatian words *lûk* (En. bow) and *lùk* (En. onion) differ by the type of the accent. If we do not write accents in text (as mentioned before, accents are not usually written in Croatian), then *luk* (En. bow) and *luk* (En. onion) will have the same graphemic sequences and are considered homograms. If, however, we write the accents—*lûk* (En. bow) and *lùk* (En. onion)—then they will no longer have the same graphemic sequences and will therefore be considered as heterographs [26]. In heterophonic homograms, the distinctive role of prosodic units is realized in changing the grammatical and lexical meanings of the form of one word; for example, the word *roda* can be *ród̀a* (nominative, singular, feminine of the

noun *róda*, En. stork) and *ròda* (genitive, singular, masculine of the word *rôd*, En. *gender*). Then, different accents can also signify different meanings of the same homograms, in their basic forms; for example, *pas* can be *pȁs* (En. dog) and *pâs* (En. waist). Different forms of the same word can have different accents as well; for example, the word *sela* (derived form from word *selo*, En. village) can be *sèla* (genitive, singular) or *sȅla* (nominative, plural).

In some of the aforementioned examples where the words have the same written form but different meanings, different MSDs, and/or different accents, that is, they belong to different lemmas such as the homogram *roda*, *róda* (nominative, singular, feminine of the noun *róda*, En. stork) and *ròda* (genitive, singular, masculine of the word *rôd*, En. gender), the problem of ambiguity can easily be solved using the earlier described Croatian accent lexicon [6] that contains words in their basic and derived forms. Each entry in the lexicon has the word form without the accent, the corresponding MSD, and the corresponding word form with the marked accent (e.g., rôd N-msnn rod). Of course, this is not the solution to the problem of nouns having the same written form (when accents are not marked in text) and different accents but the same MSDs, such as *pȁs*—MSD: N-msn- (En. dog) and *pâs*—MSD: N-msn- (En. waist). To solve this problem, morphosyntactic analysis is not enough, but in order to resolve the ambiguity, a semantic analysis should also be conducted since the problem can only be solved with the help of the context.

From our perspective of resolving the ambiguity with the help of the lexicon where each entry has the form of the word without the marked accent with its corresponding MSD and the form with the marked accent, it is important to stress out the possible relationships that the two ambiguous words can have regarding the MSD and the type of the accent.

So, there can be homograms with different MSDs and the same MSD. Within each of the two groups, pairs can have the same accents or different accents. The possible relationships between homograms are shown in Fig. 1.

V. Disambiguation Procedure

In this paper, we experimented with homograms of the category of nouns of different gender. In our future work, we plan to experiment with homograms within other categories of words (verbs, adjectives, etc.) and also across all categories since there are, for example, homograms which can belong to two different categories, such as in the word *skup*, *skùp* noun (En. gathering) and *skûp* adjective (En. expensive).

The greatest problem in detecting the cases of homograms, that is, two or more words that have the same written form but distinctive meaning, is that the lists of homograms for Croatian do not exist. There are several examples in some papers like [26], but there are no complete lists of homograms. So, one of our goals in this paper and in our future work is to make such lists that can be used in detecting ambiguous words in texts. This could prove useful, for example, in MT. If we can detect the homograms in source-language texts, we will be able to pay attention to such cases in the preprocessing procedure or in the training phase.

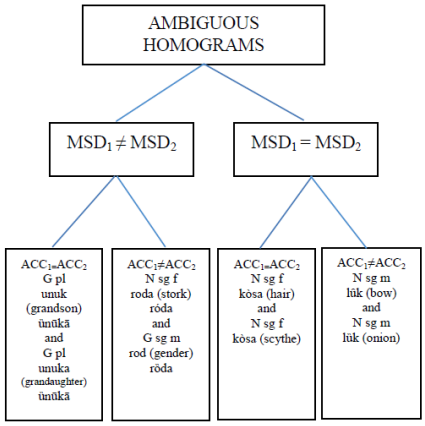

Fig. 1. Relationships between ambiguous homograms regarding the MSD and accent types.

In this paper, we limit ourselves to distinguishing the homograms of the category of nouns of different gender. In order to get the list of all such cases, we used the above-mentioned Croatian accent lexicon [6]. We searched through all nouns in the lexicon and looked for the entries that have the same written form in the word form without a marked accent but different MSD; specifically, we searched for such pairs of the same written form but of different gender. When we distinguished all such pairs, we found a set of homograms (there are no duplicate entries) and sorted them alphabetically. We added a lemma with the corresponding accent to each of the entries. The meaning can easily be disambiguated since different lemmas refer to different distinctive meanings. For example, regarding the aforementioned example of the word *roda*, the entries are shown in Table I.

TABLE I: Entries in the Lexicon of Homograms for the Word "*Roda*"

Homogram	Accent	MSD	Lemma
roda	ródā	N-fpg-	róda
roda	ròda	N-msg-	rôd

This way, we obtained the lexicon of homograms which, so far, contains nouns of different gender. The lexicon can be a valuable resource in detecting the possible ambiguous words in texts. If we detect the word *roda* in the text as a homogram (we can easily check the occurrences of the homograms listed in the lexicon), this can be a signal that we have to pay some further attention to it and check which meaning the homogram holds; we can determine the MSD of the homogram, check in the lexicon of homograms what the corresponding lemma is (with its corresponding accent), and accordingly disambiguate the meaning of the homogram. In our future work, we plan to add the sense/meaning for each entry and thus further simplify the process of disambiguation.

The obtained lexicon[1] contains a total of 19,366 entries in the form as shown in Table I. There are 3,460 unique homograms detected (a set of entries in the first column in the lexicon). This means that there are on average 5.6 entries

[1]We made it publicly available at http://elematic.uniri.hr/resources under the CC BY-NC-SA 4.0 license. If you use it, please cite this paper.

per homogram in a lexicon. This happens because there are some words that can have more than one MSD for the same form. For example, the word *vrȁta* (En. door) can have the MSDs N-npg-, N-npn, and N-npa-. The word *vrâta* (genitive of the word vrat, En. neck) has the MSD N-msg-. So, in this case, the homogram *vrata* can have four MSDs. Homograms are mainly of masculine (10,287) and feminine (8,121) gender, and only some are of neutral (958) gender. There are altogether 8,029 nouns in plural and 11,337 in singular number. Regarding the cases, the greatest number of nouns is in the vocative case (3,416), then in nominative (3,249), accusative (3,116), genitive (2,891), dative (2,535), locative (2,535), and instrumental (1,624). The complete statistics of the lexicon are shown in Table II.

TABLE II: STATISTICS OF THE OBTAINED LEXICON OF HOMOGRAMS

	Entries	19,366
	Unique homograms	3,460
	Lemmas	3,121
Gender	Masculine	10,287
	Feminine	8,121
	Neutral	958
Number	Singular	11,337
	Plural	8,029
Case	Nominative	3,249
	Genitive	2,891
	Dative	2,535
	Accusative	3,116
	Vocative	3,416
	Locative	2,535
	Instrumental	1,624

A. Identifying and Disambiguating Homograms in Corpus

We wanted to check how many homograms that we identified in the above described lexicon can be found in the available corpora for the Croatian language. In the Croatian National Corpus 3.0 (CNC 3.0) [27] which contains about 234,000,000 tokens, we found altogether 2,908,182 occurrences of the identified homograms and 2,488 occurrences of unique homograms, which make up 72% of all homograms that we obtained in the lexicon. The Croatian web corpus hrWaC [28] is a web corpus that contains 1.9 billion tokens and is annotated with the lemma, morphosyntax, and dependency syntax layers. There were 20,694,430 total occurrences of the homograms and 3,180 of unique homograms (92% of the homograms from the lexicon) identified in the hrWaC corpus. Since in our future work we want to further explore the role of the disambiguation of homograms in MT for Croatian–English language pair, we also wanted to see how many homograms we can expect in parallel corpora: SETimes [29] and hrenWaC [30]. Since those corpora are much smaller than the previous two, it is reasonable to expect that there will be a smaller number of occurrences of homograms. We identified a total of 62,462 occurrences of homograms in SETimes and 39,938 in hrenWaC. There were 949 (27% of all homograms in the lexicon) unique homograms found in SETimes and 1,221 (35% of all homograms in the lexicon) in hrenWaC. Detailed information about the representativeness of homograms in different corpora is shown in Table III.

We also chose ten homograms that were among the most frequent in the corpus and randomly chose ten sentences for each homogram in which they occur. Table IV contains details about the chosen homograms sorted by frequency in a descending order.

TABLE III: REPRESENTATIVENESS OF HOMOGRAMS IN AVAILABLE CORPORA

	Different homograms	Total occurrences	Percentage of homograms found in corpus	Percentage of corpora
CNC 3.0	2,488	2,908,182	72%	1.3%
hrWaC	3,180	20,694,430	92%	1.7%
SETimes	949	62.462	27%	1.6%
hrenWaC	1.221	39.938	35%	1.7%

TABLE IV: HOMOGRAMS USED IN THE EXPERIMENT

Homogram	MSD	Accent	Lemma	Meaning	No. of occurrences in corpus
rata	N-msg-	rȁta	rȁt	War	39,805
rata	N-fsn-	râta	râta	Installment	
vrata	N-npg- N-npn- N-npa-	vrȁta	vrȁta	Door	18,603
vrata	N-msg-	vrâta	vrât	Neck	
uloga	N-msg- N-mpg-	úloga	úlog	Investment	8892
uloga	N-fsn- N-fpg-	ùloga	ùloga	Role	
supruga	N-fsn-	sùpruga	sùpruga	Wife	7,431
supruga	N-msg-	sùpruga	sùprug	Husband	
šuma	N-fpg- N-fsn-	šûmā šùma	šùma	Forest	5,601
šuma	N-msg-	šûma	šûm	Noise	
vezu	N-fsa-	vèzu	vèza	Connection, relationship	4,637
vezu	N-msl-	vézu	vêz	Berth	
učenici	N-fsd- N-fsl-	ùčenici	ùčenica	Schoolgirl	4,910
učenici	N-mpn-	ùčenici	ùčenik	Student	
roda	N-fsn-	róda	róda	Stork	2,051
roda	N-msg-	rôda	rôd	Gender, humanity	
bitka	N-msg-	bìtka	bítak	Essence	1,402
bitka	N-fsn-	bìtka	bìtka	Battle	
kostima	N-msg-	kostíma	kòstim	Costume	1,035
kostima	N-fpd- N-fpl-	kóstima	kôst	Bone	

We applied the above described procedure of disambiguation, and this way all homograms in all sentences were successfully disambiguated. An example is given in Table V.

TABLE V: EXAMPLE OF HOMOGRAM DISAMBIGUATION IN SENTENCES

Croatian	Kad su oko **vrata [N-msg- vrâta vrât]** dobili zlatne medalje, nestalo je sve.
English	*When they got gold medals around their **neck**, everything disappeared.*
Croatian	Došao je i pokucao na naša **vrata [N-npa- vráta vráta]**.
English	*He came and knocked on our **door**.*

VI. HOMOGRAMS IN MT

Some lexical items can refer to more than one concept. This is the attribute of all homograms: they all have at least two distinctive meanings. In MT, the right meaning of a homogram needs to be chosen. It is a rather easy task for humans (under the assumption that they speak fluently both source and target languages) because they possess "the knowledge of the world." However, it is not an easy task for a machine. Wrong choices in MT systems often result in major translation mistakes, as one homogram often refers to two or more completely unrelated concepts. We assessed the performance of MT systems for the Croatian–English language pair to check whether the disambiguation of homograms could improve the performance of MT systems in avoiding major translation mistakes related to assigning the wrong meaning to homograms.

TABLE VI: GT AND MICROSOFT TRANSLATOR TRANSLATIONS OF HOMOGRAMS USED IN THE EXPERIMENT

Croatian	*Ovo sliči objavi trgovinskog **rata**, koji će teško pogoditi naše izvoznike.*
GT	This is similar to a trade **war**, which will hardly hit our exporters.
Microsoft Translator	This looks like publish a trade **war**, which will be difficult to guess our exporters.
Croatian	*Sada funkcionari Hajduka trebaju otputovati u Španjolsku i dogovoriti dinamiku isplate, a prva **rata** bi trebala, kako su obećali Španjolci, doći vrlo brzo.*
GT	Now the Hajduk officials have to go to Spain and agree on the payout dynamics, and the first **~~war~~** should, as the Spaniards promised, come very quickly.
Microsoft Translator	Now officers travel to Spain, the need of Hajduk and arrange the dynamics of war, and the first **payments** should, as they had promised the Spaniards, coming very soon.
Croatian	*Vrijednost **uloga** denominiranih u nacionalnim valutama zemalja članica Monetarne unije zbog uvođenja eura neće se promijeniti.*
GT	The value of the **~~roles~~** denominated in the national currencies of the member states of the Monetary Union due to the introduction of the euro will not change.
Microsoft Translator	The value of the **~~role~~** of the denominated in the national currencies of the Member countries of the Monetary Union because of the introduction of the euro will not change.
Croatian	*Moja prva **uloga** u životu bila je upravo na Igrama - bio sam Vlaho u Dundu Maroju.*
GT	My first **role** in life was exactly on the Game - I was Vlaho in Dundo Maroja.
Microsoft Translator	My first **role** in life she was just at the games - I was Blaise in Dundu Maroju.
Croatian	*Kad su oko **vrata** dobili zlatne medalje, nestalo je sve.*
GT	When the gold medal was received around the **~~door~~**, everything went missing.
Microsoft Translator	When they are around the **~~doors~~** get gold medals, it's gone all.
Croatian	*Došao je i pokucao na naša **vrata**.*
GT	He came and knocked on our **door**.
Microsoft Translator	He came and knocked on our **door**.
Croatian	*Predsjednika su u bolnici posjećivali njegova **supruga** i djeca s obiteljima kao i njegovi suradnici.*
GT	The president was visited by his **wife** and children with families as well as his associates.
Microsoft Translator	Of the President in the hospital visited by his **wife** and children with their families as well as his associates.
Croatian	*Vašeg **supruga** i vas ne vežu samo obiteljske veze.*
GT	Your **~~wife~~** and you do not only have family ties.
Microsoft Translator	Your **husband** and you not just made family ties.
Croatian	*Primjer je sječa tropskih **šuma** u Brazilu, u porječju Amazone.*
GT	An example is the logging of tropical **forests** in Brazil, Amazone.
Microsoft Translator	An example is the felling of tropical **forests** in Brazil, in the Amazon basin.
Croatian	*Osim tankog sloja života na samoj površini Zemlje, povremene neustrašive letjelice i ponešto radio **šuma**, naš je utjecaj na svemir ravan ništici.*
GT	Apart from the thin layer of life on the very surface of the Earth, occasional fearless aircraft and some radio **noise**, our influence on the universe is quite null and void.
Microsoft Translator	Except for a thin layer of life on the surface of the Earth, the occasional intrepid aircraft and some radio **~~forest~~**, our influence on the universe even zero.
Croatian	*Zbog toga je bila u bolnici, a **vezu** je pokušala prekinuti, što joj nije uspjelo.*
GT	That's why she was at the hospital and tried to break the **connection**, which she did not succeed.
Microsoft Translator	That is why he was in the hospital, and the **connection** is attempted to stop, which she failed.
Croatian	*Nakon cijele noći dežuranja na **vezu**, gdje su nam pukla dva privezna konopa, kad je svanulo prebacili smo se na sigurniji vez uz jednu kočaricu.*
GT	After the entire night of the **~~connection~~**, where two mooring ropes were fired, when we were late we moved to a safer berth with one bunker.
Microsoft Translator	After a whole night of dežuranja on a **~~link~~**, where we snapped two mooring rope, when the day came, we switched to a safer berth with one kočaricu.
Croatian	*Punu torbu tih podmetača dali su našoj **učenici** da unese u školu.*
GT	The full bag of these pads gave our **~~students~~** to enter the school.
Microsoft Translator	A whole bag full of these pads they gave our **~~students~~** to enter the school.
Croatian	*Po njezinim riječima, **učenici** su vrlo zainteresirani i motivirani za takav praktičan i iskustven način učenja demokracije.*
GT	In her words, the **students** are very interested and motivated for such a practical and experiential way of learning democracy.
Microsoft Translator	Per her words, the **students** are very interested in and motivated for such a practical and iskustven way of learning democracy.
Croatian	*U Čigoču obitava samo stotinjak stanovnika, ali i tristotinjak **roda**.*
GT	In Čigoč there are only a hundred inhabitants, but also a hundred and thirty **~~generations~~**.
Microsoft Translator	In Čigoču lives only about a hundred residents, but also about three hundred of the **~~genus~~**.

Croatian	*Možemo li što učiniti, makar i u posljednjemu hipu, da okrenemo novi list u povijesti ljudskoga **roda**?*
GT	Can we do, even in the last hip, to turn a new leaf in the history of **humanity**?
Microsoft Translator	Anything we can do, even if it's in the posljednjemu a second, to turn a new leaf in the history of the human **race**?
Croatian	*No, u stvarnosti u predizbornoj kampanji vodit će se oštra **bitka** između Laburista i konzervativaca.*
GT	But in reality in the pre-election campaign there will be a tough **battle** between Labor and Conservatives.
Microsoft Translator	However, in reality, the election campaign will lead to a harsh **battle** between Labor and the conservatives.
Croatian	*Cipra, naime, stvari postavlja problemski, pa knjigu dijeli na poglavlja o problemu **bitka**, o ontološkoj diferenciji bitka i bića, o hijerarhiji bića itd.*
GT	Cyprus, in fact, places things problematic, and shares the book with chapters on the problem of ~~**battle**~~, the ontological difference between the battle and the beings, the hierarchy of the beings,
Microsoft Translator	Of Cyprus, namely, the stuff sets in order, so the book is divided into chapters about the problem of ~~**battle**~~, about the ontological diferenciji battle and creatures, about the hierarchy of beings, etc.
Croatian	*Tijelo biva zakopano, ali sve dok još ima mesa na **kostima**, duša luta.*
GT	The body is buried, but as long as there is still meat on the **bones**, the soul wanders.
Microsoft Translator	The body is buried, but until there's more meat on her **bones**, soul wanders.
Croatian	*Promjene glumca često zahtijevaju i promjenu **kostima**, čak promjenu materijala?*
GT	Actors' changes often require changing the ~~**bones**~~, even changing the material?
Microsoft Translator	The actor changes often require and change **costumes**, even the change of the material?

Besides, by assessing the performance of state-of-the-art MT for the Croatian–English language pair, we might get an insight into whether it is safe to assume that integrating extra linguistic information such as accents into MT would improve the translation results for that particular language direction.

We conducted a manual evaluation of ambiguous homograms and their suggested translations. We opted for a manual evaluation to make sure that the MT system is not "punished" for producing a synonym or a paraphrase. An artificial test set of sentence pairs with ambiguous homograms is created by selecting two sentences from the Croatian National Corpus per each homogram given in Table IV. Although sentences were randomly selected, special care was taken that two senses of the selected homograms are included. One sentence serves as an example of usage in one sense, and the other sentence serves as an example of usage in another sense. Google Translate (GT) engine[2] and Microsoft Translator[3] are used for performing translations. We fed GT and Microsoft Translator with original sentences and recorded translation outputs. A complete list of sentence pairs with their respective GT and Microsoft Translator translations is given in Table VI.

We report the number of disambiguation errors in total. Croatian homograms are italicized and highlighted in bold, correctly translated words are highlighted in bold, and strikethrough is used to indicate wrongly translated words.

Out of 20 sentences (each sentence contains one homogram) that we used in MT, the GT system correctly translated 55% of the homograms, and the Microsoft Translator correctly translated 65% of the homograms. In most of the cases, one sense in a pair of senses in each of the homograms was translated correctly and the other was translated incorrectly. We presume that this happens because one sense of the homogram occurs more often in corpora than the other. This is, for example, the case with the homograms *uloga*, *vrata*, *vezu*, and *bitka*.

In most of the cases, both MT systems made the same mistakes: they both translated correctly one sense of the homogram and incorrectly the other sense. There are, however, differences with some homograms, for example, with *rata*, *supruga*, and *kostima*, which were translated correctly in both senses by Microsoft Translator and correctly in one sense but not the other by GT. On the other hand, the opposite happened with the homogram *šuma*, which was correctly translated in both senses by GT and in one sense but not the other in Microsoft Translator.

Both MT systems made a mistake when translating one sense of the homogram *roda* (N-fsn-, En. stork), and they translated correctly the other sense, *roda* (N-mng-, En. gender, sex, genus, kind, race, kindred, etc.). The complexity in the translation of this homogram is manifested in the polysemy of the corresponding translation equivalent in English. This can lead us to another point in resolving ambiguities, which might be worth paying further attention in MT.

Since both MT systems made major translation mistakes when translating homograms by assigning completely wrong meanings to them (GT in 45% of the cases and Microsoft Translator in 35% of the cases), it is safe to assume that there are cases in which homogram disambiguation could clearly improve MT.

VII. Conclusion

This paper described the procedure that we used to create a lexicon of homograms that is comprised of all Croatian nouns of different gender, which have the same written forms (in texts without marked accents) but different meanings, MSDs, and possibly different accents. The lexicon consists of 19,366 entries and 3,460 unique homograms. Each entry in the lexicon comprises the homogram, the corresponding MSD, the accented word, and the accented lemma. The experiments showed that the identification and disambiguation of homograms within the corpus with the help of the obtained lexicon are efficient and accurate. Therefore, we plan to expand the lexicon with the homograms within other categories of words (verbs, adjectives, etc.) and also across all categories, since there are examples of homograms that can belong to two different categories.

We also evaluated the performance of MT systems for the Croatian–English language pair in order to check whether the ambiguity of homograms affects the performance and whether there are possibilities of improving the performance of MT systems by disambiguation of homograms. More specifically, we

[2]Available at https://translate.google.com/.
[3]Available at https://www.bing.com/translator.

wanted to see whether the disambiguation of homograms can assist MT systems in avoiding major translation mistakes related to assigning the wrong meaning to homograms.

Since MT systems made major translation mistakes when translating homograms because of their ambiguity, we can conclude that we can improve the performance of MT by employing disambiguation of homograms.

As our future work, we plan to redesign the experiment so as to directly use accents as a preprocessing step to MT as well as to include homograms as a factor to the phrase-based translation or as additional input features into the NMT model. This would provide better comparison grounds.

REFERENCES

[1] Z. Jelaska, *Fonološki Opisi Hrvatskoga Jezika: Glasovi, Slogovi, Naglasci*, Zagreb: Hrvatska sveučilišna naklada, 2004.

[2] M. Yip, *Tone*, Cambridge University Press, 2002.

[3] E. Pletikos, "Akustički opis hrvatske prozodije riječi," Ph.D. dissertation, University of Zagreb, 2008.

[4] B. W. Fortson IV, *Indo-European Language and Culture*, Blackwell Publishing, 2005.

[5] F. Kortlandt, "The origin of the Japanese and korean accent systems," *Acta Linguistica Hafniensia*, no. 26, pp. 57-65, 1993.

[6] L. N. Prskalo, "Automatsko predviđanje i modeliranje hrvatskih prozodijskih obilježja na temelju teksta," Ph.D. dissertation, University of Zagreb, 2016.

[7] Š. Vintar, D. Fišer, and A. Vrščaj, "Were the clocks striking or surprising? Using WSD to improve MT," in *Proc. the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012, pp. 87-92.

[8] R. Sennrich and B. Haddow, "Linguistic Input Features Improve Neural Machine Translation," in *Proc. the First Conference on Machine Translation*, Berlin, Germany, 2016, pp. 83-91.

[9] A. Rios, L. Mascarell, and R. Sennrich, "Improving word sense disambiguation in neural machine translation," in *Proc. the Conference on Machine Translation (WMT)*, vol. 1, Copenhagen, Denmark, 2017, pp. 11-19.

Z. Zhong and H. T. Ng, "Word sense disambiguation improves information retrieval," in *Proc. the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 2012, pp. 273-282.

[11] M. Hadni, S. El Alaoui, and A. Lachkar, "Word sense disambiguation for arabic text," *The International Arab Journal of Information Technology*, vol. 13, no. 1A, pp. 215-222, 2016.

[12] D. Vickery, L. Biewald, M. Teyssier, and D. Koller, "Word-sense disambiguation for machine translation," in *HLT '05 Proc. the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, 2005, pp. 771-778.

[13] P. Resnik and D. Yarowsky, "Distinguishing systems and distinguishing senses: New evaluation methods for Word Sense Disambiguation," *Natural Language Engineering*, vol. 5, no. 2, pp. 113-133, 1999.

[14] P. Koehn and H. Hoang, "Factored translation models," in *Proc. the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, 2007, pp. 868-876.

[15] R. Sudarikov, O. Dušek, M. Holub, O. Bojar, and V. Križ, "Verb sense disambiguation in machine translation," in *Proc. the Sixth Workshop on Hybrid Approaches to Translation*, Osaka, Japan, 2016, pp. 42-50.

[16] N. Bakarić, J. Njavro, and N. Ljubešić, "What makes sense?" *InFuture 2007*, Zagreb, 2007.

[17] D. Alagić and J. Šnajder, "Experiments on active learnin for croatian word sense disambiguation," in *Proc. the 5th Workshop on Balto-Slavic Natural Language Processing*, Hissar, Bulgaria, 2015, pp. 49-58.

[18] D. Alagić and J. Šnajder, "Cro36WSD: A lexical sample for croatian word sense disambiguation," in *Proc. the 10th Language Resources and Evaluation Conference (LREC 2016)*, Portoroz, Slovenija, 2016.

[19] E. Barić *et al.*, *Hrvatska gramatika*. Zagreb: Školoska knjiga, 1995.

[20] S. Babić *et al.*, *Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika*. Zagreb: Globus, Nakladni zavod, 1991.

[21] S. Vukušić, I. Zoričić, and M. Grasselli-Vukušić, *Naglasak u hrvatskome književnom jeziku*, Zagreb: Nakladni zavod globus, 2007.

[22] N. Mikelić Preradović, "Pristupi izradi strojnog tezaurusa za hrvatski jezik," Ph.D. dissertation, University of Zagreb, 2008.

[23] V. Anić, *Veliki rječnik hrvatskoga jezika*.: Novi liber, 2009.

[24] N. Ljubešić, *MULTEXT-East Morphosyntactic Specifications*, Croatian Specifications, Apr. 29, 2013.

[25] T. Erjavec *et al.*, "The MULTEXT-east morphosyntactic specifications for slavic languages," in *Proc. the EACL 2003 Workshop on the Morphological Processing of Slavic Languages*, Budimpešta, 2003, pp. 25-32.

[26] B. Tafra, "Istopisnice i istoslovnice u hrvatskom jeziku," in *Slavenski jezici u usporedbi s hrvatskim II*, D. Sesar, Ed. Zagreb: FF press, Faculty of Humanities and Social Sciences, University of Zagreb, 2011.

[27] M. Tadić, "New version of the croatian national corpus," in *After Half a Century of Slavonic Natural Language Processing*, D. Hlaváčková *et al.*, Eds. Brno: Masaryk University, 2009.

[28] N. Ljubešić and F. Klubička, " {bs,hr,sr}WaC - Web corpora of Bosnian, Croatian and Serbian" in *Proc. the 9th Web as Corpus Workshop (WaC-9)*, Gothenburg, 2014.

[29] Ž. Agić and N. Ljubešić, "The SETimes.HR linguistically annotated corpus of croatian," in *Proc. the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, 2014.

[30] N. Ljubešić, D. Berović, D. Merkler, and M. Tadić, *hrevWac*, 2013.

Lucia Nacinovic Prskalo was born in 1983. She earned a PhD degree in information and communication sciences at the Faculty of Humanities and Social Sciences, Zagreb, Croatia in 2016. Her major field of study is natural language processing.

She is a postdoctoral researcher at the Department of Informatics of the University of Rijeka in Rijeka, Croatia. Her research interests include natural language processing, speech synthesis, speech and language technologies, speech prosody, machine learning, text classification, and text mining.

Marija Brkic Bakaric was born in 1983. She earned a PhD degree in information and communication sciences at the Faculty of Humanities and Social Sciences, Zagreb, Croatia in 2013. Her major field of study is machine translation.

She is an assistant professor at the Department of Informatics of the University of Rijeka in Rijeka, Croatia. Her research interests include data mining, text mining, machine learning, and natural language processing.