THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

*PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,*
*Volume 15, Number 1/2014, pp. 68–75*

# ROMANIAN-ENGLISH SPEECH TRANSLATION

Tiberiu BOROŞ, Dan TUFIŞ

"Mihai Drăgănescu" Research Institute for Artificial Intelligence, Romanian Academy
Corresponding author: Tiberiu BOROŞ, E-mail: `tibi@racai.ro`

Speech to speech (S2S) translation is a complex process designed to enable the communication between individuals that speak different languages and it represents a valuable contribution to (1) science, (2) cross-cultural interaction and (3) global business. Through S2S, a text spoken in one language is automatically recognized, translated and synthesized in another language. This paper presents an overview of our approach to Romanian-English bi-directional speech translation and we cover the methods and technologies used for implementing such a system

*Key words*: automatic speech recognition, machine translation, speech synthesis, speech to speech translation.

## INTRODUCTION

Recent technological advances and breakthroughs leading to the constant increase in computational power with an unprecedented tendency of building smaller and smaller devices that yield higher and higher performance, has inevitably lead to a strong demand for information retrieving [1] and communication enabling technologies that are multilingual aware. Speech-to-speech (S2S) translation is a complex process designed to assist communication between individuals that speak different languages. Through S2S, a text spoken in one language is automatically recognized, translated and synthesized in another language.

While in the past it was impossible to link technologies such as ASR, TTS and MT in a single portable device due to hardware constraints, recent technological advances and breakthroughs have opened new horizons for assistive technologies based on human-computer interaction. The computational power of current smart-phones and tablets is orders of magnitude higher than those of the desktop computers of the early 2000's. Internet access has become more a requirement than a luxury, enabling computational tasks on mobile devices that in the past were either impossible or prohibitively expensive.

There are several projects and systems designed for S2S translation [2, 3, 4] mainly centered on Arabic-English, Japanese-English and Chinese-English language pairs. In this paper we present an overview of our approach to Romanian-English bi-directional speech translation and we brief on the methods and technologies used for implementing such a system. Because of its complex nature, speech translation presents a series of challenges. In order to provide an optimal input for MT, (1) spellchecking and (2) diacritic normalization and restoration has to be performed on the output of the ASR component. Our research has shown that the best translation results are produced by (3) a cascaded translation using (4) factored translation model which involves (5) part of speech tagging and (6) lemmatization and surface translation model. Finally, synthesizing arbitrary text, requires additional text processing such as (7) letter-to-sound conversion, (8) syllabification and (9) lexical stress prediction. All the above mentioned processing requirements are plagued by the issue of data-sparseness and the presence of out-of-vocabulary (OOV) words which require special attention.

## 1. SYSTEM OVERVIEW

A speech translation system requires three basic components: a multilingual ASR component capable of delivering accurate results in real world noise environment, a MT component that is suited to accurately perform bi-directional translation between the languages of interest and a multilingual TTS system capable of synthesizing natural, pleasant sounding and intelligible voices. While this is a straightforward and explicit requirement, research has shown that the independent design of these components lacks joint optimality [5], and by coupling ASR with MT the system performs generally better. Tying together these systems is not a trivial task. While the MT and TTS systems are developed entirely in-house in our approach, we resorted to an external ASR solution: Google Speech Recognition. The Google ASR API was chosen because it offers the necessary support for robust speech recognition under noisy environments, it does not require any speaker adaptation steps and it is already adapted for mobile devices. Google Speech Recognition uses a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) with a triphone model [6]. Our informal evaluation included multiple speakers both in a quiet office and in an outside-noisy environment and it showed that the system has a word accuracy rate ranging from 92% to 96% depending on the speaker and environment. However, the recognition results from Google ASR are not directly usable within our speech translation system (see Fig. 1 and further details in Section 4). Because we already presented parts of the MT and TTS systems in previous works, we will only focus on the challenges posed by the current speech translation task, namely the data sparseness issue for highly inflectional languages (discussed in chapter 3) and the ASR post-editing (discussed in chapter 4) while only providing a brief review of the modules.

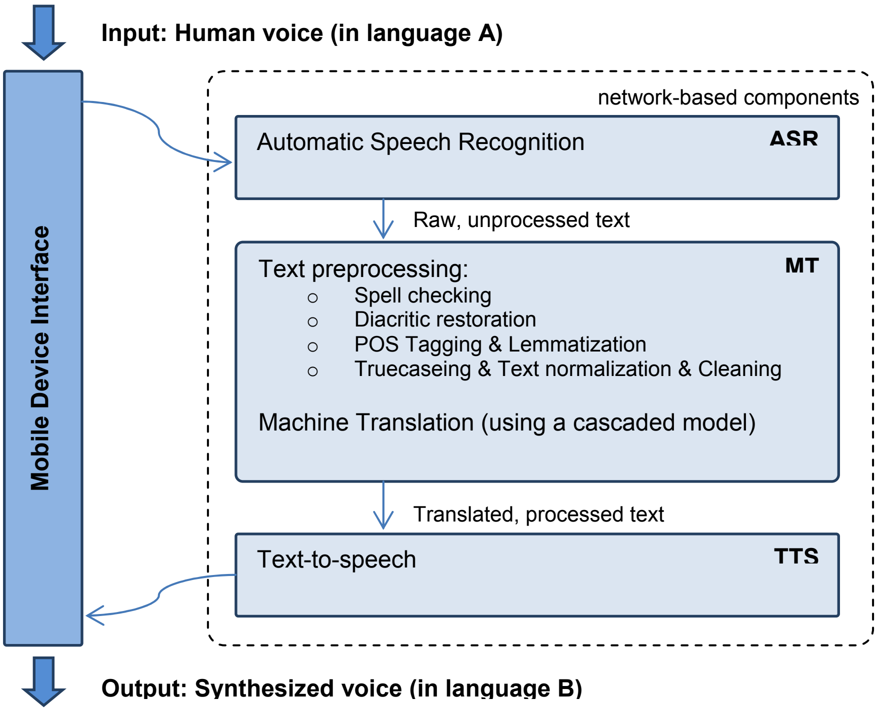
Input: Human voice (in language A)
Mobile Device Interface
network-based components
Automatic Speech Recognition
ASR
Raw, unprocessed text
Text preprocessing:
MT
Spell checking
Diacritic restoration
POS Tagging & Lemmatization
Truecaseing & Text normalization & Cleaning
Machine Translation (using a cascaded model)
Translated, processed text
Text-to-speech
TTS
Output: Synthesized voice (in language B)

Fig. 1 – System architecture diagram.

To prove the viability of our concept we implemented the above mentioned platform on an Android mobile platform. The storage requirements to make this platform standalone are expensive, as it would be required to include the models for lemmatization, POS tagging, machine translation, and speech synthesis (tens of Gb). Furthermore, the Google's ASR solution is only available as a Web service, thus we preferred to make the entire prototype network-based and access the required services for MT and TTS remotely. This way, the prototype can be extended to cover multiple languages, exploiting our already existing SMT systems for Romanian, English, German, Spanish [7] and French [8].

## 2. NATURAL LANGUAGE PROCESSING AND THE ISSUE OF DATA SPARSENESS

The vast majority of Natural Language Processing (NLP) tasks are based on statistical methods and machine learning techniques. For a language with rich morphology the number of inflected forms of a dictionary headword may be large enough, thus generating data sparseness issues for a data-driven approach. That is, during the learning phase, occurrences of all possible inflectional forms should be seen in the training corpus. If different word-forms are treated independently without including any higher level linguistic knowledge, the corpora requirements for creating a translation model is prohibitively expensive and unfeasible. Factored translation models extend the phrase based translation by taking into account not only the surface form of the phrase, but also additional information like the dictionary form (lemma), the part-of-speech tag or the morpho-syntactic specification. They also provide, on the target side, the possibility to add a generation step. All these new features accommodate well in the log-linear model employed by many decoders:

$$P(e \mid f) = \mathrm{e\,x\,p} \sum_{i=1}^{n} \lambda_i h_i(e, f), \tag{1}$$

where $h_i(e, f)$ is a feature function associated with the pair *(e, f)* and $\lambda_i$ is the weight of the function.

## 2.1. Part-of-speech tagging

Factored translation is designed to reduce the effect of data sparseness for highly inflectional languages and as such it requires key text-processing steps such as lemmatization and part-of-speech (POS) tagging. According to the Multext-East lexical specifications [9], Romanian requires a number of approximately 1200 lexical tags, also referred to as morpho-syntactic descriptors (MSDs). By exploiting the language specific syncretism, the number of MSDs was reduced to 614 different tags. These MSDs encode part-of-speech information with associated attributes inside a string, in which each attribute has a pre-defined position. The first character is an upper case character denoting the part of speech (e.g. ‘N’ for nouns, ‘V’ for verbs, ‘A’ for adjectives, etc.) and the following characters (lower letters or ‘-‘) specify the individual lexical attributes of the specified POS. For example, the MSD ‘Ncfsrn’, specifies a noun (the first character is ‘N’) the type of which is common (‘c’, the second character), feminine gender (‘f’), singular number (‘s’), in nominative/accusative case (‘r’) and indefinite form (‘n’). Non-relevant attributes for a language, or for a given combination of feature-values are marked using the character ‘-’. The trailing hyphens are omitted. For a language which does not morphologically mark the gender and definiteness features, the earlier exemplified MSD will be encoded as ‘Nc-sr’. The MSD set is too large for a tagger standard training procedure, with a real data sparseness threat. There are several methodologies designed to address the data sparseness issue for POS tagging such as Tiered Tagging [10], [11] or the Neural MSD Tagger [12]. The

Tiered Tagging methodology explores the reduction of the MSD tagset by removing context-recoverable attributes. The newly obtained reduced tagset (six times smaller) is called a CTAG set and the tagging procedure is two-step: a standard tagger assigns CTAGs to the words inside an utterance and then, post-processing by rule-based or ML techniques are employed in order to extend CTAGs to MSDs based on information from the local context of the word/CTAG.

The Neural MSD Tagger uses a neural network to iteratively assign tags from left to right to words inside an utterance. For each word, the tagger makes its decision based on features constructed from the tags assigned to preceding words and the features of the possible tags assignable to the succeeding words. The network requires an encoding method for converting the MSD strings to real-valued feature vectors which are actually used in the training and tagging process. The encoding algorithm explores the fact the attribute types are not always unique to their category in order to reduce the encoding space. The MSDs are numerically encoded into feature vectors, the size of the feature vectors depending on the global number of attributes of all possible parts of speech. The vectors either represent fixed tags (for previously tagged words inside an utterance) or possible tags (for words that have not been processed yet) that are computed based on using the maximum-likelihood estimation (MLE). Table 1 shows the encoding of the Romanian MSD ‘Rw-n’, belonging to the Adverb category.

*Table 1*

Example of encoding for MSD ‘Rw-n’

| Category | | | | | | | | | | | | | | | | Clitic | | | Degree | | | | Type-Adverb | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | **7** | 8 | 9 | 10 | 11 | 12 | 13 | 14 | … | 41 | **42** | … … | 61 | 62 | 63 | … … | 98 | 99 | 100 | 101 | **102** | 103 | … … | 139 |
| *J* | *N* | *V* | *A* | *P* | *D* | *T* | ***R*** | *S* | *C* | *M* | *Q* | *I* | *Y* | *X* | | *y* | ***n*** | | *p* | *c* | *s* | | *g* | *p* | *z* | *m* | ***w*** | *c* | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |

Currently, the Neural MSD Tagger is based on a 50 neuron hidden layer network with a custom topology obtained using an unrestricted genetic algorithm, and ensures an average accuracy of more than 98% [13].

## 2.2. Lemmatization

Lemmatization is the process of determining a word's canonical form (lemma or dictionary entry) from its inflectional form. It is a technique useful in various natural language processing applications such as data-mining and document classification [14] and as previously mentioned it is used by factored translation models to reduce the size of translation tables. In the case of English, the lemmatization process is fairly simple, but for highly inflectional languages, such as Romanian, this process poses a series of challenges. In our approach we used a Perceptron classifier trained by the Margin Infused Relaxed Algorithm (MIRA). To be able to use the MIRA framework, we had to reformulate lemmatization as a sequence labelling task. As such, each individual letter of a word was treated as an individual token and the system was designed to assign labels. Each label assigned to a token was used to denote one of the following transformations:

– '*' – means leave current letter *unchanged*

– '*_nil_*' – means that the current letter must be *removed* from the word's lemma

– '*_r(<character sequence>)* – means that the current letter has to be *replaced* with the character sequence in brackets (*<character sequence>*).

As an example, the labeling for the lemmatization of the inflected form "brazi" (English "firs") which has the canonical form "brad" (fir) is the following: b/* r/* a/* z/d i/_nil_. Note the d/z consonant alternation.

The lemmatization process has to take into account the information provided by the word's morpho-syntactic-description (MSD) tag. This means that we either have to train different models for different MSDs or we have to incorporate the MSD information inside the features we use. The Romanian MSDs inventory is very large even after exploring the syncretism (more than 600 MSDs) and consequently, the model obtained by training with MSDs is extremely large, difficult to train and use. In order to reduce our lemmatization model size, we converted every word's MSD from our training set into a CTAG, based on the Tiered Tagging methodology, and we included only the inflectional open grammatical categories nouns, verbs and adjectives. This reduced our model size about 5 times.

The context used by the labeler is composed of both *lexical* and *morpho-syntactic* features (CTAGs): $(l_{-2},l_{-1},l,C)$, $(l_{-3},l_{-2},l_{-1},l,C)$, $(l_{-4},l_{-3},l_{-2},l_{-1},l,C)$, $(l,l_1,l_2,C)$, $(l,l_1,l_2,l_3,C)$, $(l,l_1,l_2,l_3,l_4,C)$, $(l_{-1},l,l_1,C)$, $(l_{-2},l_{-1},l,l_1,l_2,C)$, where $l$ is used to mark the current letter, $l_i$ is used to denote the letter at relative distance $i$ from the current one and $C$ is used to denote the word form's CTAG. Using a word-form lexicon composed of 1.2 M words we withheld 10% for each individual CTAG as the test set. The results of our experiments are shown in Table 2. The overall accuracy was **94.19%,** which is **12%** higher than the results presented in [15].

*Table 2*

Experimental results with lemmatization

| CTAG | # of tokens | # of errors | Accuracy % | CTAG | # of tokens | # of errors | Accuracy % |
|---|---|---|---|---|---|---|---|
| **A** | 16 | 0 | 100 | **V2** | 8195 | 664 | 91.9 |
| **VN** | 871 | 47 | 94.6 | **NPOY** | 6427 | 1092 | 83,01 |
| **NSON** | 4223 | 190 | 95.5 | **V3** | 7312 | 629 | 91.4 |
| **APOY** | 5078 | 99 | 98.05 | **ASON** | 3030 | 43 | 98.58 |
| **NSVN** | 79 | 3 | 96.2 | **VPPM** | 797 | 58 | 92.72 |
| **ASN** | 6205 | 65 | 98.95 | **NSRY** | 6701 | 104 | 98.45 |
| **VPSM** | 1178 | 77 | 93.46 | **VPPF** | 747 | 15 | 97.99 |
| **NSOY** | 6761 | 279 | 95.87 | **V1** | 6180 | 455 | 92.64 |
| **ASRY** | 5121 | 67 | 98.69 | **APRY** | 5119 | 95 | 98.14 |
| **NP** | 263 | 35 | 86.69 | **NSRN** | 4244 | 19 | 99.55 |
| **NPRY** | 6443 | 884 | 86.28 | **ASOY** | 5122 | 59 | 98.85 |
| **VG** | 2973 | 118 | 96.03 | **NPN** | 6615 | 1223 | 81.51 |
| **NN** | 263 | 3 | 98.86 | **NPVY** | 28 | 3 | 89.29 |
| **VPSF** | 748 | 15 | 97.99 | **NSVY** | 2225 | 31 | 98.61 |
| **APN** | 6062 | 127 | 97.9 | **ASVY** | 626 | 12 | 98.08 |
| **NSN** | 2591 | 6 | 99.77 | **AN** | 106 | 6 | 94.34 |
| **Overall** | | | | | **112349** | **6523** | **94.19** |

In Table 2, all CTAGS beginning with an "N" are nouns, "A" are adjectives and "V" are verbs. The best result (100%) is for invariant adjectives ("A") for which the lemma is the word form. This behaviour is preserved for all CTAGs for which lemma is equal to the word form: NSRN (noun, singular, nominative/accusative, non-definite form) with 99.5%, ASN (adjective, singular, non-definite form) with 98.95%, etc.  We analysed the lemmatization errors and discovered that the gold-standard lexicon contained a few hundreds of erroneous lemmatized entries. The major source of real lemmatization errors was in the vast majority of cases rare nouns (NPRY, NPOY and NPN) either neologisms or regionalisms. As previously mentioned, this evaluation was performed without having access to the lemmas stored in the lexicon. In regular running conditions, the entire information (including lemmas of the more than 1.2 million word-forms stored in the lexicon) is available to the NLP processor and given that from our experience the average number of out-of-vocabulary words in an arbitrary text was never higher than 6-7% of the total number of words in a text, we estimate that the number of errors in the processing chain, due to wrong lemmatization, is negligible (less than 0.6%).

## 3. MACHINE TRANSLATION ON AUTOMATIC SPEECH RECOGNITION DATA

The text produced by the Google ASR interface needs further preprocessing for usage within our machine translation system. When input is Romanian, the ASR result requires cleaning up, diacritic restoration and normalization. For example, the utterance "1 2 3" is recognized by the ASR system as "1doi3". By performing a simple search on Google, one can find that "1doi3.ro" is a website and the result produced by the ASR system is somewhat expected, since it is primarily designed for speech recognition of search queries. Also, most recognized text does not include diacritics and when diacritics are used, the system uses the old-style convention characters for 'ş' and 'ţ' (*s-cedil* instead of *s-comma* and *t-cedil* instead of *t-comma*) so, they need normalization. When input to the ASR is English, the result needs spell-checking (especially when the speaker is non-native in English). Another very useful kind of normalization is called truecasing. This process means lower-casing the first word in every sentence, where necessary and using the upper case letters for acronyms or proper nouns. A truecase model is trained on available target language data and it benefits automatic machine translation when building the translation model and the language model by reducing the number of surface forms for each possible word, directly addressing the issue of data sparseness and thus allowing for better machine translation performance. In our experiments [8], module of truecasing contributed translation quality improvements in the range of 1-2 points of BLEU scores [16].

### 3.1. The RACAI Spellchecker

The RACAI spellchecker for Romanian and English is designed to produce alternative spellings with a decision threshold to replace words inside an utterance with their corrected form. It is a corpora-based method which was thoroughly presented in [17], combining 3 algorithms for spellchecking using a voting mechanism. All algorithms use similar approaches to spellchecking: (1) Detect if a word is correctly spelled using dictionaries; (2) If the word is not found in any lexicon, produce spelling alternatives by deleting, replacing or adding letters and spaces; (3) Re-rank spelling alternatives for the entire utterance using n-gram frequencies (see equation 2).

The system was tested for English within the Microsoft Speller Challenge Competition (placing 4th) and obtained an F-score of 97% on the TREC DATASET [18].

$$S(q_i) = \left(\alpha\sum P(w_i) + \beta\sum P(w_i, w_{i+1}) + \gamma\sum P(w_i, w_{i+1}, w_{i+2})\right)F(q, q_i) \tag{2}$$

$\alpha, \beta, \gamma$ – weights

$P(w_i)$, $P(w_i, w_{i+1})$ and $P(w_i, w_{i+1}, w_{i+2})$ – n-grams log probabilities

$F(q,q_i)$ – factor dependent on Levenshtein distance between the original query $q$ and the spelling variant $q_i$.

### 3.2. Diacritic restoration

Diacritic restoration is one type of spelling correction in which the correct diacritical mark of a letter is inserted in a word which would otherwise be incorrect, have a different (unintended) meaning or violate different syntactic constraints for the language in question. We use the DIAC+ system [19], specifically designed for Romanian, which inserts the diacritics based on the context of the word and it differentiates among the following cases:

1. The word is incorrect according to a predefined (large) lexicon but a diacritic version of it exists in the lexicon, e.g. “mașina” is correct, “masina” is not;
2. The word does not possess the correct diacritic form to agree with its syntactic constraints, e.g. the indefinite noun in “o mamă” (“a mother”) is correct but its definite form is not “o mama” (“the a mother”);
3. The word does not have the intended meaning in context, e.g. the word “fata” means “the girl” but word “fața” means “the face”.

In Romanian [19] the morpho-syntactic information obtained by POS tagging the diacritic-free text is, for the vast majority of cases, sufficient to solve the ambiguities that occur when deciding whether to introduce a diacritic or not. For instance, the sequence “o mam**a**” is tagged with an indefinite article and an indefinite noun but the only correct form for “mam**a**” when it is an indefinite noun is “mam**ă**”. In Romanian (cf. [19]), on average, every third word of an arbitrary text contains at least one diacritical character. In terms of characters, more than 8.2% have diacritical signs. Depending on how one evaluates, the accuracy of DIAC+ is 99.4% when all the characters in a text are considered and 95.1% when only the characters that require diacritics are counted.

### 3.3. Cascaded translation model

There are multiple approaches to machine translation that fall within one of the three classes: rule-based, statistical and hybrid. Statistical based methods are the prevalent approaches for implementing machine translation systems today. The phrase based translation approach has overcome several drawbacks of the word based translation methods and proved to significantly improve the quality of translated output. The morphology of a highly inflected language permits a flexible word order, thus shifting the focus from long range reordering to the correct selection of a morphological variant. Morphologically rich languages have a large number of surface forms in the lexicon to compensate for a flexible word order. Both Transfer and Interlingua MT employ a generation step to produce the surface form from a given context and a lemma of the word. In order to allow the same type of flexibility in using the morpho-syntactic information in translation, factored translation models [20] provide the possibility to integrate the linguistic information into the phrase based translation model. Most of the statistical machine translation (SMT) approaches that have a morphologically rich language as target employ factored translation models.

In our approach, the effective translation is done by a cascaded translation model [21] using a first layer factored translation model (S1) and a second layer surface translation model (S2). The hypothesis is that by training a second phrase-based statistical MT system (S2) on the data that was output by our initial system (S1), this second system will correct some of the errors the initial system made. This approach was experimentally validated in [21] showing a 0.39 BLEU point increase for the Romanian to English translation direction on the TED free-speech genre text corpus. The major advantage of the cascaded translation model is that the second translation system does not need additional data, being trained on the same data that the first system was trained upon. There is however one downside to this approach: the translation time of a sentence will be doubled as the sentence will pass through two distinct systems instead of only one. But, considering that to translate a sentence takes on the order of tens to a few hundred milliseconds on current hardware, cascading two systems for real-time translation is acceptable.

## 4. TEXT-TO-SPEECH SYNTHESIS

TTS synthesis is an extremely important process in improving accessibility by enabling voice controlled systems to interact with users. In TTS an input text is converted into spoken language while undergoing a series of complex tasks such as POS tagging, letter-to-sound (LTS) conversion, syllabification, lexical stress prediction and so on. The TTS component of a speech translation system handles the conversion of the text output of the MT component into spoken language. The context of bi-directional speech translation implies that the TTS component has to be *multilingually oriented* as well as synthesizing intelligible, natural and pleasant voices.

Significant effort has been invested in trying to improve the naturalness of the synthesized voice and to increase the level of acceptance of TTS systems among the users. Still, the main difference between TTS systems and other systems that allow computer-human interaction using spoken language (e.g. interactive voice response (IVR) systems using pre-recorded sentences) is that a TTS system must be able to synthesize voice starting from arbitrary text. The quality of the synthetic voice is influenced by the numerous factors: the text pre-processing steps (e.g. part-of-speech tagging, homograph disambiguation, syllabification, lexical stress prediction, letter-to-sound conversion etc.), the speech synthesis method used by the system (e.g. concatenative, statistical parametric) and the size and quality of the speech corpora on which the system was trained. The key to obtaining high quality voices lies in either (1) working with narrow domains, (2) using large-scale speech corpora or (3) resorting to statistical parametric voices.

The system developed at RACAI [22] uses state-of-art methods for all the lexical and morphological processing steps involved in TTS and is a multilingual oriented platform allowing rapid prototyping. It implements the Perceptron with Margin Infused Relaxed Algorithm (MIRA) training for sequence labelling, combining various methods such as the onset-nucleus-coda (ONC) encoding for syllabification [23] (99% accuracy for Romanian on OOV words), EM alignments with sequence labelling for L2S conversion [24] (96.29% for Romanian on OOV words) and also original approaches to lemmatization (more than 94% accuracy on OOV words) and lexical stress prediction [25] (98.80%). Furthermore, the previously presented Neural MSD tagger used by our system enables easy adaptation to other highly inflectional languages.

To test the multilingual capabilities of our TTS system, RACAI's system entered, in early 2013, the international Blizzard Challenge 2013 with a system for English built in less than 3 weeks. It performed more than reasonably, ranking 7th among the 14 competing systems, out of which the first three were state-of-the-art commercial ones. Our participation in the Blizzard Evaluation Challenge was rewarding because:

a) A synthetic voice for English was built, proving that the data-driven methods implemented by our system can easily be adapted to work with languages other than Romanian;
b) The system was optimized for large scale speech corpora, enabling us to build high quality natural voices based on the very large training data provided by the organizers;
c) The system was modified to work with the stylistic annotation provided in the contest, showing that our system is easily adaptable;
d) Text from different genres such as news and audio books was processed.
e) The several dimensions which the evaluations were carried on identified strong and weak points of our system, giving hints for further developments.

## 5. CONCLUSIONS AND FUTURE WORK

We presented recent work in our NLP group at RACAI, on speech translation, a very complex task involving automatic speech recognition, machine translation and text-to-speech synthesis. We reviewed the problems posed by the data-sparseness of highly inflectional and morphologically rich languages describing our solutions for each of the sub-processes involved.

Currently the SMT platform incorporates prototype translation systems for several language pairs trained using open parallel corpora (JRC Acquis[1], DGT-TM[2], Europarl[3], OPUS[4] etc.) or automatically extracted by means of the LEXACC Tool [27] from Wikipedia. These data sources do not fall in the same genre as normal spoken language. Future development plans are aimed at improving our system by creating additional resources that are better fit for the task of speech translation and improving our TTS system by collecting and annotating additional speech corpora. Because network traffic is generally a limiting factor on

[1] http://optima.jrc.it/Acquis/index_2.2.html
[2] http://ipsc.jrc.ec.europa.eu/index.php?id=197
[3] http://www.statmt.org/europarl/
[4] http://opus.lingfil.uu.se/

most mobile data-plans and sending voice data over the Internet is more demanding than sending plain text, we will experiment with a reduced TTS model that can be embedded directly into the device, thus providing an alternative to the networked speech synthesis module.

## REFERENCES

1. ION, R., ŞTEFĂNESCU, D., & CEAUŞU, A., *Important Practical Aspects of an Open-domain QA System Prototyping*. Proceedings of the Romanian Academy, Series A, **9**, *3*, 253–258, 2008.
2. NEY, H., *Speech translation: Coupling of recognition and translation*, Acoustics, Speech, and Signal Processing, IEEE International Conference on., Vol. 1, 1999, pp. 517–520.
3. ZHANG, R., KIKUI, G., YAMAMOTO, H., WATANABE, T., SOONG, F., & LO, W. K., *A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation*, Proceedings of the 20th COLING (*Conference on Computational Linguistics*), Association for Computational Linguistics, Geneva, 2004, pp. 1168–1174.
4. WAIBEL, A., AHMED BDRAN, A., BLACK, A.W., FREDERKING, R.E., GATES, D., LAVIE, A., LEVIN, L.S., LENZO, K., TOMOKIYO, L.M., REICHERT, J., SCHULTZ, T., WALLACE, D., WOSZCZYNA, M., ZHANG, J., *Speechalator: two-way speech-to-speech translation on a consumer PDA*., Proceedings of Interspeech, 2003.
5. LAVIE, A., WAIBEL, A., LEVIN, L., FINKE, M., GATES, D., GAVALDA, M., & ZHAN, P., JANUS-III: *Speech-to-speech translation in multiple languages*, Acoustics, Speech, and Signal Processing (ICASSP-97), IEEE International Conference on., Vol. 1, 1997, pp. 99–102.
6. JAITLY, N., NGUYEN, P., SENIOR, A. W., & VANHOUCKE, V., *Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition*, Proceedings of INTERSPEECH, 2012.
7. TODIRAŞCU, A., ION, R., NAVLEA, M., & LONGO, L., *French text preprocessing with TTL*, Proceedings of the Romanian Academy, Series A, **12**, *2*, pp. 151–158, 2011.
8. TUFIŞ, D., ION, R., DUMITRESCU, S. AND STEFĂNESCU, D., *Wikipedia as an SMT Training Corpus*, Proceedings of the 9th conference RANLP*,* Hissar, Bulgaria, September 10–13, 2013.
9. ERJAVEC, T., & MONACHINI, M., *Specifications and notation for lexicon encoding*, COP Project, 106, 1997.
10. TUFIŞ, D., *Tiered tagging and combined language models classifiers. In Text, Speech and Dialogue*, Springer, Berlin, Heidelberg, 1999, pp. 28–33.
11. TUFIŞ, D., & DRAGOMIRESCU, L., *Tiered tagging revisited*, Proceedings of the 4th LREC Conference, 2004, pp. 39–42.
12. BOROS, T., ION, R., TUFIS, D., *Large tagset labeling using Feed Forward Neural Networks. Case study on Romanian Language*, Proceedings of the 51st Conference of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, August 3–7, 2013.
13. BOROŞ, T., DUMITRESCU, Ş.D., *Improving the RACAI Neural MSD tagger*, Proceedings of the 14th International Conference on *Engineering Applications of Neural Networks* (EANN 2013), Halkidiki, Greece, September 13–16, 2013.
14. TUFIŞ, D., POPESCU, C., & ROŞU, R., *Automatic classification of documents by random sampling,* Proceedings of the Romanian Academy, Series A., **1**, *2*, pp. 18–28, 2000.
15. ION, R., *Word Sense Disambiguation Methods Applied to English and Romanian* (in Romanian), PhD Thesis Romanian Academy, Bucharest, 2007.
16. PAPINENI, K., ROUKOS, S., WARD, T., ZHU, W. J., BLEU: *A method for automatic evaluation of machine translation*, Proceedings of the 40th ACL Conference: *Annual meeting of the Association for Computational Linguistics,* 2002, pp. 311–318
17. ŞTEFĂNESCU, D., ION, R., BOROŞ, T., *TiradeAI: An Ensemble of Spellcheckers*, Proceedings of the Spelling Alteration for Web Search Workshop*,* 2011, pp. 20–23.
18. QIN, T., LIU, T. Y., & LI, H., *The TREC Datasets in LETOR*, Part of the TREC datasets in LETOR description, 2007.
19. TUFIŞ, D., & CEAUŞU, A., *DIAC+: A professional diacritics recovering system*, Proceedings of the 6th Conference LREC, 2008.
20. KOEHN, P., & HOANG, H., *Factored Translation Models*, EMNLP-CoNLL, 2007, pp. 868–876.
21. TUFIŞ, D., DUMITRESCU, S. D., *Cascaded Phrase-Based Statistical Machine Translation Systems,* Proceedings of the 16th Conference of the European Association for Machine Translation, 2012, pp. 129–136.
22. BOROS, T., STEFĂNESCU, D., ION, R., *Handling Two Difficult Challenges for Text-to-Speech Synthesis Systems: Out-of-Vocabulary Words and Prosody – A Case Study in Romanian*, in A. Neustein, J.A. Markowitz (eds.), *Where Humans Meet Machines*, Springer, 2013, pp. 137–163.
23. BARTLETT, S., KONDRAK, G., CHERRY, C., *Automatic syllabification with structured SVMs for letter-to-phoneme conversion*, Proceedings of ACL-08: HLT, 2008, pp. 568–576.
24. JIAMPOJAMARN, S., CHERRY, C., & KONDRAK, G., *Joint processing and discriminative training for letter-to-phoneme conversion*. Proceedings of 46th Conference ACL: HLT, 2008, pp. 905–913.
25. BOROŞ, T., *A unified lexical processing framework based on the Margin Infused Relaxed Algorithm. A case study on the Romanian Language*, Proceedings of The 9th Conference RANLP*,* Hissar, Bulgaria, September 10–13, 2013.
26. ŞTEFĂNESCU, D., ION, R., HUNSICKER, S., *Hybrid parallel sentence mining from compara-ble corpora,* Proceedings of the 16th Conference of the European Association for Machine Translation, Trento, Italy, 2012.

*Received July 24, 2013*