Statistical Machine Translation of Broadcast News from Spanish to Portuguese

Raquel Sánchez Martínez[1], João Paulo da Silva Neto[2],
and Diamantino António Caseiro[1]

L²F - Spoken Language Systems Laboratory, INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
{raquel.sanchez,dcaseiro}@l2f.inesc-id.pt,
Joao.Neto@inesc-id.pt
http://www.l2f.inesc-id.pt/

Abstract. In this paper we describe the work carried out to develop an automatic system for translation of broadcast news from Spanish to Portuguese. Two challenging topics of speech and language processing were involved: Automatic Speech Recognition (ASR) of the Spanish News and Statistical Machine Translation (SMT) of the results to the Portuguese language. ASR of broadcast news is based on the AUDIMUS.MEDIA system, a hybrid ANN/HMM system with multiple stream decoding. A 22.08% Word Error Rate (WER) was achieved in a Spanish Broadcast News task, which is comparable to other international state of the art systems. Parallel normalized texts from European Parliament database were used to train the SMT system from Spanish to Portuguese. Preliminary non-exhaustive human evaluation showed a fluency of 3.74 and sufficiency of 4.23.

Keywords: Automatic Speech Recognition, Broadcast News Transcription, Acoustic Model, Language Model and Statistical Machine Translation.

1 Introduction

One of the main motivations beyond this research work was the opportunity to expand an existing and optimized Portuguese broadcast news recognition system to process Spanish broadcast news context and consequently to calculate it performance in different languages domain. In the best of our knowledge, is the first broadcast news machine translation system for the Spanish to Portuguese language pair, what did an appealing target.

A great focus has been placed in ASR research area due to emerging demands, for example, from people with hearing disabilities. This have driven an elevated research level and generated a great variety of services and commercial applications. Technological advances in recent years as digital signal processors, faster and affordable memories and increased capacity have also contributed in the evolution of ASR system.

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 112–121, 2008.

The SMT research has regain focus research, after a few years where it was left aside in favor of linguistic knowledge representation, mainly due to not be comparable the results and the necessary effort for the latter area.

The costs involved in manual translation by a professional translator have also driven the companies to use SMT as an attractive solution.

A large amount of data is available for English Broadcast News enabling the development of concurrent ASR systems for this task. The current state-of-art WER is less than 16% [1] in real-time (RT) operations and under 13% [2] with 10 times RT.
There are Spanish Broadcast News Recognition Systems based on reference English Recognition Systems developed by research centers CMU [3] and BBN [4] [5], co-financed by DARPA, and IBM, LIMSI and RWTH within a project co-founded by the European commission [6].

For the development of the Spanish system the data available at LDC[1] was used in a total of 30 hours of Latin America Broadcast News audio and different Spanish newspapers corpus.

Table 1. Vocabulary and corpora text dimension, in number of words (Mw: Million words; Kw: Thousands of words) and the respective WER for the different system

	LIMSI_04	IBM_04	RWTH_04	BBN_97	CMU_97	BBN_98
TOTAL(Mw)	400	210	140	157	157	157
VOCABULARY(Kw)	65	47	50	40	40	40
WER(%)	17.8	23.3	17.8	19.9	23.3	21.5

The table 1 presents the total number of words contained in text corpora, the vocabulary size used to develop the language model and the WER of each system. RWTH_2004 has the best WER of 17.8% [6]. CMU_1997 and IBM_2004 have 23.3% WER [3] [6] which are the worst values in the systems under study. These values were obtained with the test set “1997 Hub-4 BN Evaluation Non-English Test Material” by LDC. It corresponds to one-hour Latin America Broadcast News audio, with the same acoustic conditions as those used in the acoustic model training.

The work involved in the development of Statistical Machine Translation of Broadcast News from Spanish to Portuguese has started with the study of a platform that was already being applied to Portuguese Broadcast News task [7]. Then audio and text were selected in order to create a lexicon, acoustic models and language models for the Spanish recognition system. This system was evaluated showing comparable results to other international state-of-the-art systems. Parallel normalized texts of both languages were used to train the translation probabilities and to develop the SMT system.

This paper is organized in the following way: in the section 2 it is described the selection and transformation process of necessary corpora; section 3 explains the changes made in the existing Portuguese Recognizer to adapt it to Spanish and in section 4 it is presented the translation system from Spanish to Portuguese. The last section present the conclusions and future work.

[1] http://www.ldc.upenn.edu/

2 Corpora Description

In order to allow the comparison of this work with the studied state-of-the-art systems, the same corpora sources were applied, whenever possible.

2.1 Audio Corpora

We use 30 hours of Latin America Broadcast News audio made available by the LDC, which represents mostly Cuba and Mexico dialect. Their corresponding transcriptions were normalized and transformed to AUDIMUS.MEDIA [7] [8] system.

The acoustic files are divided in 80 NIST SPHERE format files, without compression. The data are 16-bit linear PCM, 16-KHz sample frequency, single channel. Most files contain 30 minutes of recorded material and some contain 60 or 120 minutes. The sampling format requires roughly 2 megabytes (MB) per minute of recording, so the file sizes are typically around 60 MB, with some files ranging up to 120 or 240 MB. The transcripts are in SGML format, using the same markup conventions.

This corpus is divided in 23 hours, corresponding to 63 files, for training set (75%), 4 hours, 10 files, for development set (15%) and 3 hours, 7 files, for test set (10%). The audio selection process for building each set tried to give similar coverage to each phone in the different dialects, creating a more robust acoustic model to the dialectal variability. On the other hand, news of older dates were used in training set and news of more recent dates were used in development and test set, avoiding the context-dependence between news of near dates in each different set.

2.2 Text Corpora

A statistical language model requires a large quantity of data that should be adapted to the task to obtain proper probabilities. We had available a corpus with audio corpora training set transcriptions, they are totally adapted to Latin America Broadcast News task, but with a total of 300,000 words, being an insufficient dimension to generate a statistical language model. We created another corpus using a newspapers set from LDC, namely “Spanish Gigaword First corpus Edition”, adding newspapers from previous editions of “English News Text” and “Spanish Newswire Text”, which were not included in the last edition. They constitute a large data set of about 720 Million words, necessary to generate the statistic language model, despite the newspapers grammatical constructions are more formal than in Broadcast News. Text corpus was divided in training (75%), development (15%) and test (10%) sets, following the same rules than audio corpora. It was normalized by removing labels, punctuation and special symbols. Other normalization step expanded abbreviations, numbers and acronyms.

2.3 Parallel Text Corpora

The parallel text corpus consists of proceedings of the European Parliament session. This corpus was assembled by Philipp Koehn [9] and has been extensively used by researchers in Statistical Machine Translation. The language used in this corpus is more formal than Broadcast News, and consists of approximately 1.3 Million sentence pairs.

It was necessary to normalize this parallel text, deleting formatting tags and punctuation, and expanding abbreviations. For the Spanish side, we used the same tools as used for preparing the language model corpus. For the Portuguese side, an existent normalizer tool was used. Finally, we removed sentences deemed too long or non-existent in one of the languages, obtaining approximately 700,000 sentence pairs to train the SMT system.

3 Spanish Broadcasts News Recognizer

3.1 Introduction

The automatic broadcast news recognition is still a challenge due to not resolved questions, since the almost frequent and unpredictable changes, in the speaker, type of speech, the topic, vocabulary, and the record and channel conditions, between others. Then a very important work in this research area is the obtaining of big quantities of audio and text resources with these characteristics included. Language model, acoustic model, vocabulary and lexicon to Spanish task did not exist previously to our work, being necessary to develop a complete system with specific language tools.

3.2 Reference Platform

The AUDIMUS.MEDIA system [7] is a hybrid speech recognition system that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs).

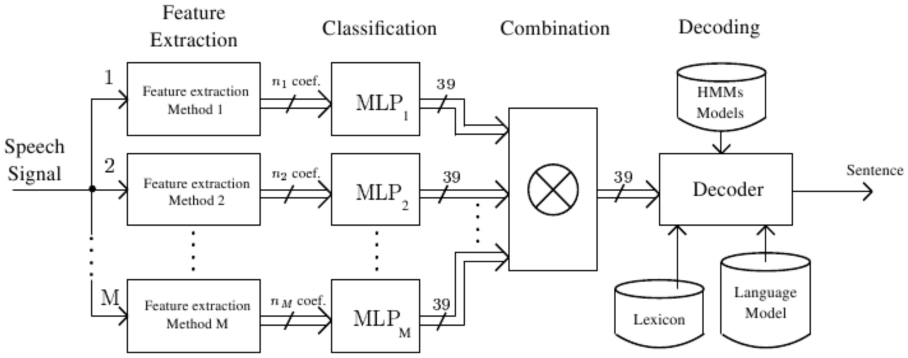

Fig. 1. AUDIMUS.MEDIA recognition system and processing stages [7]

In figure 1 is represented AUDIMUS recognition system. The first stage of processing comprises speech signal feature extraction where three different methods are used PLP [10], log RASTA [10] and MSG [11]. In the MLPs classification stage posteriori probabilities are generated to the 39 possible context-independent phones, 38 for the Portuguese language, and 1 representing the silence. The individual MLP processing results are combined in the next stage.

It use Weighted Finite-State Transducers Technology (WFST) [12] in the decoding stage, where combined phone stream is processed to produce the resulting sentences based on a finite vocabulary and a language model that represents the task restrictions set.

3.3 Vocabulary and Lexical Model

A similar dimension of vocabulary compared with state-of-the-art systems was desired. First we selected 70,500 most common words from the newspapers corpus. Then, we added some not included words from transcriptions training set. Finally, we filtered foreign words and uncommon acronyms to obtain 64,198 words vocabulary.

It was used an automatic grapheme to phone transcription system similar to [13] to generate the phonetic transcriptions. The lexicon uses symbolic representation from SAMPA, plus [N] and [z] phones. The total phone set comprises 32 phones, including a silence phone. In addition, we made manual lexicon corrections of typical foreign words. It was also created, a program based on regular expressions that detects abbreviations and transcribes them using a set of rules.

3.4 Alignment and Training of Acoustic Model

We use generic acoustic model without speaker-dependence, due to the high number of speakers in the corpus audio training set. In order to create this acoustic model, it is necessary an initial phone/audio alignment.

One of the options available was to train the model with small Spanish audio corpora with good acoustic conditions, as in CMU [3]. However it was decided to obtain the initial Spanish acoustic model by transformation of the Portuguese Broadcast News optimized model as in LIMSI [6], because of the difficulty to generate manual alignments at phone level, necessary in the first option.

We generate a phone mapping between the 23 Spanish phones (22 sound phones, plus a silence phone), and the 39 Portuguese phone set. For the remaining 9 Spanish phones there is not direct mapping. To solve this problem, we chose the Portuguese phone set with the most similar sound of the Spanish phone. Having the phone transformed, and parameters optimized, the acoustic model was trained applying an iterative process.

Firstly acoustic training set was aligned by the model optimized for Portuguese Broadcast News, and then we transformed the corresponding results with mapping phones described above, obtaining initial Spain Broadcast News targets. We made 4 alignments and MLP training with the new Spanish network. Since after these alignments was no significant change in the recognition results, we stopped the training process.

3.5 Language Model

In order to create language model we used the transcriptions and newspapers text corpora. It was not possible to separately use them, since individually did not gather the necessary characteristics to create a robust language model. The transcription text corpus is adapted to the Broadcasts News task, but it has a small dimension (300,000 words), and Newspaper corpus has a big dimension (700 Millions words) but their form is not a good representation of spontaneous speech of Broadcast News. After studying several alternatives [14], it was decided to first generate a language model for each individual corpus with the same vocabulary and then interpolate them, reducing the perplexity that models have separately.

Table 2. Newspaper Corpora LM Perplexity and dimension for different cut-off

	n-gram					
Cut-off	1	2	3	4	Total	PPL
50-50-50	64,198	109,880	128,942	135,008	438,028	305.78
25-25-25	64,198	216,904	288,024	302,562	871,688	250.59
4-4-4	64,198	1,378,934	2,819,181	3,153,520	6,993,616	158.76
2-4-4	64,198	2,849,752	2,819,181	3,153,520	8,886,651	153.27
2-3-4	64,198	2,849,750	3,994,811	3,153,520	10,062,279	150.93
2-2-3	64,198	2,849,750	7,599,741	5,025,741	15,539,430	146.48

Table 3. Transcription Corpora LM Perplexity and dimension for 3-4-grams

	n-gram					
	1	2	3	4	Total	PPL
order 3	64,198	99,883	168,284	----------	332,365	327.53
order 4	64,198	99,883	168,284	184,690	517,055	359.07

Table 4. Interpolated LM perplexity and dimension

	n-gram					
Cut-off	1	2	3	4	Total	PPL
50-50-50	64,198	172,306	267,548	303,937	807,989	170.47
25-25-25	64,198	266,685	413,675	464,146	5,968,687	151.77
4-4-4	64,198	1,400,967	2,903,685	3,289,301	7,658,151	110.09
2-4-4	64,198	2,865,844	2,903,685	3,289,301	9,123,028	106.86
2-3-4	64,198	2,865,844	4,073,443	3,289,301	13,848,416	105.41
2-2-3	64,198	2,865,844	7,669,114	5,156,557	15,755,713	102.60

For the newspapers corpus, we generated 4-gram language models through SRLIM tools[2], using discounting of Kneser-Ney [15]. We also conducted experiences with different cut-off values [16]. Table 2 shows the perplexity (PPL) for each cut-off experiment and different n-gram orders on the development set. The obtained PPL can be considered a good representation of the adaptation of the language model to Broadcasts News task, since OOV rate is approximately 1%.

For the transcriptions corpus, we generated two N-gram language models of 3rd and 4th order. Cut-off were not apply, due to the limited corpora dimensions. The PPL results are presented in Table 3. PPL was calculated on the audio corpora development set. This value was high because of the small data size. It was chosen a 4-gram language model in order to have a greater representation in the interpolation, despite having a greater perplexity than 3-gram language model.

[2] http://www.speech.sri.com/projects/srilm/

Finally both models were interpolated with the SRILM tool, giving greater weight to the transcription language model (λ=0.68). In Table 4 is observed an improvement of perplexity in the interpolation results in relation to individual models. In the end, it was chosen a language model with cut-off of 2-3-4 and perplexity 105.41. The model with lower perplexity had a greater dimension, penalizing overall real time system's performance and increasing the write error probability in the n-gram selection.

3.6 Evaluation

Two different evaluations sets were used. WER was calculated in audio corpora test set, corresponding to 7 files and 3 hours in total. In the table 5, it is represented the different files names of the test set and their individual WER. It is observed that for su97612.sph WER is larger than the other audio files because their acoustic conditions are worse than the others. The total mean was also calculated, obtaining a value of approximately 25.62% WER.

Also the WER was calculated with the same test set (called h4ne97sp.sph, made available by the LDC) as state-of-the-art systems. In table 5 it is observed a value of 22.08% WER. This is a comparable value to those obtained in the systems studied previously.

Table 5. Test set WERs: It is represented the recognition evaluation with the test set selected from audio corpora and test set eval97 by LDC

Audio Name	WER%
se97406.sph	23.02
su97610.sph	22.16
su97610.sph	27.31
su97611.sph	27.84
su97612.sph	33.12
sv97725b.sph	20.70
sv97725c.sph	27.04
TOTAL	25.62
h4ne97sp.sph	22.08

4 Machine Translation

We decided to use a statistical approach to machine translation, as the phrase-based SMT system for Spanish to English [17]. This approach has advantages relative to others systems [18], namely, it is a language independent technology, does not require linguistic experts, allows fast creation of prototypes, and the statistical framework is compatible with the statistical techniques used in automatic speech recognition.

The corpus was based on parallel texts from European Parliament session transcriptions. The SMT system was based on Weighted Finite-State Transducers [17], and consisted of a cascade of transducers each representing the knowledge source in the SMT system, including the phrase-table and the target language model.

It was used a bootstrapping process where word-based translation models of increasing complexity and accuracy are trained and used to align each sentence pair in the corpus at the word level. This word alignment was then refined by combining automatic Spanish to Portuguese word alignment. Finally, all possible phrases were extracted from the combined alignment.

This training process was done using available tools. In particular, word level alignments used IBM 4 model [19] as implemented in GIZA ++, and the phrase-table was extracted using the MOSES[3] software package.

The phrase-table generated from the European Parliament corpus was extremely large (approximately 155 Millions phrases). In order to reduce its size, all phrases containing Spanish words not included in the speech recognition vocabulary were removed.

The resulting system was still too large for on-line use, thus an off-line system was developed which, given an input text, selects the relevant phrases from the phrase-table prior to translation. A WFST based decoder was developed for translation, which consists of a WFST representing a phrase-table. In this transducer, each simple path between the initial and final states corresponds to a particular phrase, the input labels corresponding to Spanish words and the output one to Portuguese words. Decoding is done by:

1. Converting the input sentence to a linear automaton.
2. Compose the automaton with the phrase table transducer.
3. Search the best path in the composition.

This decoder is monotonic in the sense that input and output phrases are produced in the same order, although word reordering is allowed inside each phrase. We believe that this limitation is not very important given the proximity of the two languages. Furthermore, this monotonous prevents long delays that are not desirable in a near future on-line implementation.

An initial effort to assess the translation quality of the system was done using a non-exhaustive human evaluation. Seven evaluators scored 15 translated sentences, yielding a result of 3.74 fluency and 4.23 sufficiency (in a 1 to 5 scale). These are good results in which the similarity of the two languages plays an important role.

5 Summary and Future Work

In this work we built a Statistical Machine Translation System of Broadcast News from Spanish to Portuguese. The fusion of two wide research fields was necessary.
The hybrid real-time recognition system AUDIMUS.MEDIA [7] was used as the recognition engine. After creating the acoustic models, language models, lexicon and vocabulary for the Spanish Broadcast News and carry out successive trainings, we obtained a 22.08% WER for the test eval97 to Latin America Broadcast News. This is a comparable WER value to the one produced by state-of-the-art systems, which are based on HMM models and realize several passages in the decoding stage.

The SMT strategy adapted is phrase-based translation. The MOSES software and normalized parallel texts select from European Parliament collection available were used to train the translation probabilities and models. First, a large phrases-table was

[3] http://www.statmt.org/moses/

created, and later was reduced by smaller language models adapted to the Broadcast News. In the last, there was realized a not-exhaustive human evaluation, obtaining a result of 3.74 fluency and 4.23 sufficiency.

For improvements and futures implementation we will generate Spanish and Portuguese language models with the same translated vocabulary to the SMT system and we will adapted the models to European Spanish Broadcast News.

Acknowledgments

The authors would like to acknowledge the work by Isabel Trancoso on the grapheme to phoneme transcription system to Spanish. This work was funded by PRIME National Project TECNOVOZ number 03/165.

References

[1] Matsoukas, S., Prasad, R., Laxminarayan, S., Xiang, B., Nguyen, L., Schwartz, R.: The 2004 BBN 1xRT Recognition Systems for English Broadcast News and Conversational Telephone Speech. In: Proceedings INTERSPEECH, Lisbon, Portugal (2005)

[2] Nguyen, L., Abdou, S., Afify, M., Makhoul, J., Matsoukas, S., Schwartz, R., Xiang, B., Lamel, L., Gauvain, J., Adda, G., Schwenk, H., Lefevre, F.: The 2004 BBN/LIMSI 10xRT English broadcast news transcription system. In: Proceedings INTERSPEECH, Lisbon, Portugal (2005)

[3] Huerta, J.M., Thayer, E., Ravishankar, M.K., Stern, R.: The Development of the 1997 CMU Spanish Broadcast News Transcription System. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA (1998)

[4] Kubala, F., Davenport, J., Jin, H., Liu, D., Leek, T., Matsoukas, S., Miller, D., Nguyen, L., Richardson, F., Schwartz, R., Makhoul, J.: The 1997 BBN byblos system applied to broadcast news transcription. In: Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA (1998)

[5] Matsoukas, S., Nguyen, L., Davenport, J., Billa, J., Richardson, F., Siu, M., Liu, D., Schwartz, R., Makhoul, J.: The 1998 BBN Byblos primary system applied to English and Spanish broadcast news transcription. In: Proceedings DARPA Broadcast News Workshop, Herndon, VA (1999)

[6] Westphal, M.: TC-STAR Recognition Baseline Results, TC-STAR Deliverable n<Superscript>0</Superscript> D6 (2004), http://www.tc-star.org/documents/deliverable/deliverable_updated14april05/D6.pdf

[7] Meinedo, H., Caseiro, D., Neto, J., Trancoso, I.: AUDIMUS.MEDIA - A Broadcast News speech recognition system for the European Portuguese language. In: Proceedings PROPOR, Faro, Portugal (2003)

[8] Neto, J., Martins, C., Meinedo, H., Almeida, L.: AUDIMUS - Sistema de reconhecimento de fala contínua para o Português Europeu. In: Proceedings PROPOR IV, Évora, Portugal (1999)

[9] Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005 (2005)

[10] Hermansky, H., Morgan, N., Baya, A., Kohn, P.: RASTA-PLP speech analysis technique. In: Proceedings ICASSP, San Francisco, USA (1992)

[11] Kingsbury, B.E., Morgan, N., Greenberg, S.: Robust speech recognition using the modulation spectrogram. Speech Comunication 25, 117–132 (1998)

[12] Caseiro, D., Trancoso, I.: Using Dynamic WFST Composition for Recognizing Broadcast News. In: ICSLP, Denver, CO (2002)

[13] Caseiro, D., Trancoso, I., Oliveira, L., Viana, C.: Grapheme-to-Phone Using Finite-State Transducers. In: IEEE Workshop on Speech Synthesis, Santa Monica, CA (2002)

[14] Souto, N., Meinedo, H., Neto, J.: Building language models for continuous speech recognition systems. In: Ranchhod, E., Mamede, N.J. (eds.) PorTAL 2002. LNCS (LNAI), vol. 2389. Springer, Heidelberg (2002)

[15] Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. Proceedings ICASSP 1, 181–184 (1995)

[16] Jelinek, F.: Self-organized language modeling for speech recognition. Speech Recognition 1, 450–506 (1990)

[17] Caseiro, D.: The INESC-ID Phrase-based Statistical Translation System. In: TC-STAR OpenLab, Trento, Italy (2006)

[18] Callison-Burch, C., Koehn, P.: Introduction to Statistical Machine Translation. ESSLLI Summer Course on SMT (2005)

[19] Mohri, M., Pereira, F., Riley, M.: Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language 16(1), 69–88 (2002)