TOWARDS SPEECH UNDERSTANDING ACROSS MULTIPLE LANGUAGES

T. Ward, S. Roukos, C. Neti, J. Gros, M. Epstein, S. Dharanipragada [*]

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

In this paper we describe our initial efforts in building a natural language understanding (NLU) system across multiple languages. The system allows users to switch languages seamlessly in a single session without requiring any switch in the speech recognition system. Context dependence is maintained across sentences, even when the user changes languages. Towards this end we have begun building a universal speech recognizer for English and French languages. We experiment with a universal phonology for both French and English with a novel mechanism to handle language dependent variations. Our best results so far show about 5% relative performance degradation for English relative to a unilingual English system and a 9% relative degradation in French relative to a unilingual French system. The NLU system uses the same statistical understanding algorithms for each language, making system development, maintenance, and portability vastly superior to systems built customly for each language.

1. INTRODUCTION

With the availability of commercial speech recognition systems in multiple languages, it is interesting to explore the notion of speech recognition and natural language understanding across multiple languages. For the speech recognition component, we desire a universal speech recognizer for multiple languages with a universal phonology and a mechanism to handle language dependent variations. In particular, regions like Quebec in Canada have bilingual populations conversant in both French and English. This creates the possibility of a single person conducting transactions in both languages - switching from one to another when the representation of the intent is not familiar in one of the languages. It is cumbersome for an application to have to be manually switched either through a user-interface interaction or through language ID. The language switching may occur between utterances or within a single utterance. This motivates the need for a seamless universal speech understanding system that has a "translingual capability," i.e., the system is capable of understanding user requests across multiple languages. Most current conversational systems have the following components:

- continuous speech recognition for recognizing the words that are spoken,
- context-independent natural language understanding (CI-NLU) for extracting the meaning conveyed by the recognized utterance,
- context-dependent natural language understanding (CD-NLU) for determining the meaning in the context of the conversation, and
- dialog management where the system decides on the appropriate response including appropriately generation, taking initiative, and graphical presentation.

In application domains where users may use multiple languages, an efficient system design is one where most of these components can handle the multiple languages relying on a minimum of components that are language specific. We present in this paper, our current bilingual English-French system that uses a universal phone set for speech recognition with a combined language model. For the CI-NLU components, we use statistical parsers that are trained automatically from multi-lingual corpora. The multi-lingual corpora contain natural language sentences for each target language, along with a formal representation for the meaning of the sentences (which we call the formal language). We use a single formal language for both natural languages which facilitates integrating the system into an application; the application need only process the formal language queries given to it, after the natural language understanding component has been run. For the CD-NLU component we have a unified language independent treatment of context as it is done in the formal language space. Our system allows users to switch languages between utterances while understanding their request within the context of their conversation. We present in the following sections the various components of the system.

2. METHOD

2.1. A Universal Phone alphabet

The phone sets for English and French in our previous recognizers are completely distinct. Table 1, column 1 shows the English phone set with the appropriate word context in column 2. The French phone set is shown in Table 1, column 4, with the corresponding French word in column 3. Each row in Table 1 defines the initial mapping to a universal phone set obtained with the help of human knowledge and phonetic expertise. Note that all consonants and most of the vowels are mapped to the appropriate English counterparts. This is not to say that the mapped phones are phonetically identical, rather that they are sufficiently similar for the purposes of initial acoustic modeling. We present below an algorithm that allows the data to decide whether an allophone is language-universal or language-dependent. Several phones in English do not

[*] in reverse alphabetical order

have a corresponding French phone: the whisper HH; the vowels like AH, IH, UH; the fricatives DH, JH, TH; the affricates CH and JH; and word ending stops like GD, TD, PD, etc. Phones used in the French language but not in English include nasalized vowels like AN, UN, IN and ON; the front "rounded" vowels Ê and Ě and U_; labial-palatal semi-consonant UU; and nasal velo-palatal consonant GN. The initial Universal phone set consists of the combination of common phones in Table 1, the English specific phones and French specific phones.

2.2. Language Question

In our ranks-based system [1] we build context-dependent sub-phonetic models, constructed as follows: each phone is represented by three states. A context-dependent model for each state is constructed by asking binary questions about the phonetic context P_i, for positions i= -1, -2, -3,-4,-5, for each acoustic vector aligned to the current state. Each question is of the form : is P_i in S, where S is a subset of the phonetic alphabet P. We use phonologically meaningful subsets of phones commonly used in the analyses of speech., e.g., S = p, t, k (all unvoiced stops), etc. In order to select the best question at each node from this set, an evaluation function based on a probabilistic measure related to the homogeneity of a set of parameter vectors at the node is used [2].

In contrast with other efforts to define common phone sets and acoustic models [11, 4], in this work, we experimented with a language question in addition to the phone context question while building context-dependent models using a common phone set. That is, each vector was tagged with the language identity, and a question about the language identity of the vector was added to the list of phonetic context questions. We let the data decide if the language question was a better question at each node relative to the phone context questions. The use of such a language question allows a phone-context dependent separation of acoustic models for the two languages and allows a pooling of the data for contexts in which the two languages are similar. The details of the implication of the language question and an analysis are presented elsewhere [3].

2.3. A Unified Formal Language for Natural Language Understanding

There are several methods currently used for statistical NLU. These include statistical parsers[9, 5], source channel models[8, 6], and direct channel models[7]. In each of these approaches, the systems are trained from corpora that have been annotated with the meaning. For statistical parsers, the natural language sentence is manually annotated with the correct parse tree, using a tag and non-terminal set customly designed for the application. For source channel and direct channel models, the corpus is augmented with a formal language representation. These can be semantic concepts[6] or NL-Parse statements[8]. Whether the formal language is given separately (source and direct channel models) or grafted on top of the natural language (statistical parsing), a carefully chosen formal language allows use across multiple languages. This has many benefits. First, this allows convenient bootstrapping into new languages by translating from one of the existing languages. For source and direct channel models, the for-

English Phone	English Word	French Word	French Phone
AA	CARD	adieux	A_
AO	AUSTIN	borde	Ǒ O=
AX	AGAIN	le	E=
B	BE	bien	B_
D	DEN	demi	D_
EH	YES	pre'cis	ô'
EY	SPAIN	fait, adresser	ô" ô=
F	FOR	france	F_
G	GO	augmente	G_
IY	ITALY	ligne	I_
L	LATE	ligne	L_
M	MATE	limites	M_
N	NOT	lunette	N_
OW	O	mauvais	Ô
P	PER	paris	P_
R	PROP	paris	R_
S	SO	adresse	S_
SH	RUSH	attache	CH
T	TO	retour	T_
UW	SUPER	atout	OU
V	VIA	vous	V_
W	WE	oui	W_
Y	YES	yeux	Y_
Z	ZERO	limousine	Z_
ZH	PLEASURE	alliages	J_
AE	CAN		
AH	COME		
AW	BOUND		
AXR	BEAVER		
AY	TIME		
BD	CAB		
CH	BEACH		
DD	END		
DH	OTHER		
DX	CITY		
ER	CERTAIN		
GD	LEG		
HH	HAVE		
IH	HIT		
IX	AUSTIN		
JH	JAW		
KD	KICK		
NG	KING		
OY	BOY		
PD	PROP		
TD	FLIGHT		
TH	FOURTH		
TS	STARTS		
UH	SUGAR		
		adieux	Ê
		amateur	Ě
		annule	U_
		apprend	AN
		bon	ON
		bien	IN
		emprunt	UN
		ennuis	UU
		gagne	GN

Table 1: Universal Phone Set 1: French Specific Phones

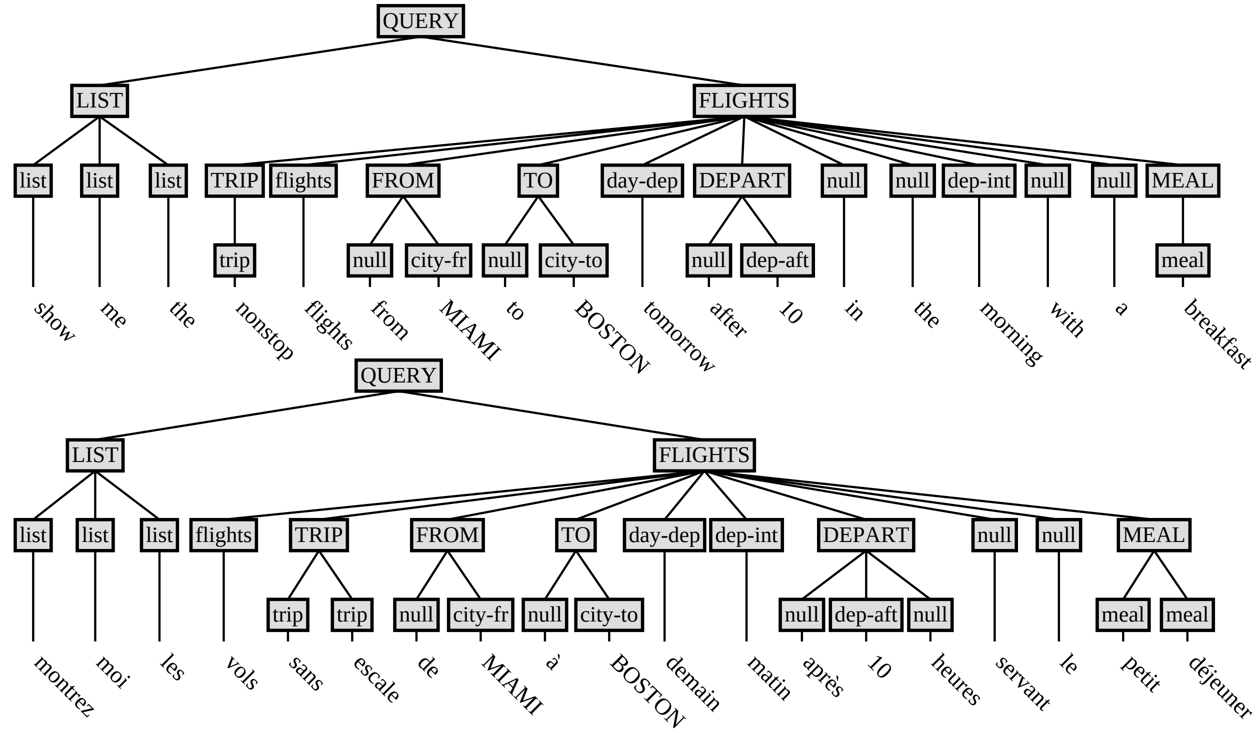

Figure 1: Example parse trees for English and French

mal language need not be modified. For statistical parsing, the new sentence must be annotated. But the annotator can be told what the nonterminals in the tree should be, speeding the annotation process. Second, a common formal language allows for a single context dependency or dialog module. Lastly, integration into a working application is much easier when the application need worry about one type of input meaning structure.

As an example, consider the English sentence:

> show me the nonstop flights from Miami to Boston tomorrow after ten in the morning with a breakfast

One French translation is:

> montrez moi les vols sans escale de Miami a Boston demain matin apres dix heures servant le petit dejeuner

The formal language used for a source or direct channel model would be:

> LIST flights nonstop FROM-CITY TO-CITY tomorrow morning DEP-AFT breakfast

The trained statistical model will learn that the formal language word nonstop likes to generate English strings like "nonstop" and "without any stops" and French strings "sans escale" and "sans arret". Likewise, notice that hand annotated parse trees (Figure 1) for these sentences use the same nonterminal nodes, even though the tree structures are different.

The statistical NLU models are trained separately for each language, as the natural language constructions used in each might be very different. For example, French adjectives generally follow the noun they modify, unlike English. Thus, after recognizing an output string, we probabilistically decide whether or not to apply the understanding using the English or French understanding models. We have not trained models where sentences contain both English and French. Though English sentences with a few French words or vice versa do not pose problems, provided examples of these are in the training data.

Once the statistical NLU has been run, the output is now language independent. So, the meaning can be interpretted in the context of the dialog, to inherit or disambiguate structures used previously in the dialog. While we have not included a multi-lingual dialog component, the multi-lingual aspect should not pose any additional problems. Our system simply switches the language used for display to match the query's language.

3. EXPERIMENTAL RESULTS

Experiments were performed on the ATIS speech recognition task. A baseline English ranks-based left-context only system [1] was trained using 16,223 training utterances from the ARPA ATIS training data and an additional 32000 read ATIS sentences collected at IBM. A class-based trigram language model was built using 16223 ARPA ATIS training utterances. A baseline pure French ranks-based left-context only system was trained using 11000 general French dictation sentences, 4000 read ATIS French sentences and 19000 read Canadian French ATIS sentences. A class-based trigram language model was built using a corpus of 16223 sentences translated from the ARPA English training sentences. The French system is built to handle the "liaison" phenomenon [10].

The English test set contained 930 utterances from the ARPA ATIS test set containing 7881 words from 27 speakers. The baseline system performance for this task was 6.3% word error (labelled "Language-specific" in Table 2). The French test set contained about 3500 words from 3 speakers. This is a smaller test set relative to the English set. The pure-French baseline performance is 12.6%. A possible reason for the higher error rate for French is the weakness of the French language model relative to English. The French Language Model was built by translating the English training sentences into French. In addition, the French data included both Parisian French and Canadian French which are somewhat different French languages.

3.1. Bilingual System

The training data for the combined system consisted of combination of English and French training sentences. The initial models for phones in "Universal Phone Set 1" were created by pooling data between French and English phones where they were mapped together and keeping the data distinct for phones that had no corresponding cross language map. Given the initial models, a ranks-based left context only system [1] was trained using the above training data. A joint class-based trigram language model was built by pooling all the English and French data. Also, a joint acoustic vocabulary containing about 6600 words was used.

The results for the initial system using **Universal Phone Set 1** are shown in row 2, Table 2. Note that a 15% degradation in English performance is seen relative to a purely English system, and a 9.5% relative degradation in French is seen with respect to a purely French system.

System	English Error rate	French Error rate
Language-specific	6.3%	12.6%
Universal Phone Set 1	7.3%	14.3%
LQ, LM penalty	6.6%	13.8%

Table 2: Recognition Error rate: Universal Phone Set 1

One of the problems with a joint model is transitions from one language to another: at some point in the search, an acoustically similar word from the second language can become the winning hypothesis if it has a reasonable unigram probability. One method to address this problem is by penalizing the transitions from language to language at the Language Model level.

We built a system that uses the trees with the **language question** using "Universal Phone Set 1" and a language transition penalty. Combining the use of a Language transition penalty with the use of a language question (LQ) (labelled LQ, LM Penalty) gave additional improvements. The English word error rate is only 5% worse relative to the pure English system, while the French word error rate is about 9% worse relative to the pure French system. Thus using a language question which allows a language-based split of the acoustic models in specific phonetic contexts, appears to be beneficial.

Natural Language Understanding: The parser for French was built from a set of French ATIS sentence that were translated from the English training set. About 6,000 French sentences were hand parsed. The resulting treebank was used to build the French parser. The same software used for building the English system was used to build the French system. The accuracy of the French parser was slightly worse than the English parser (exact match for English is around 74% while it is 64% for French.) The French treebank was smaller than the English one and was not done as carefully. The bilingual system is currently running in realtime as a web-based interface for travel information.

4. REFERENCES

1. L. Bahl, P.V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny. Robust methods for using context-dependent features and models in a continious speech recognizer. In *Proceedings of the IEEE ICASSP*, volume II, pages 632–635, Adelaide, South Australia, April 1994.
2. L. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny. Decision trees for phonological rules in continuous speech. In *Proceedings of the IEEE ICASSP*, volume 1, pages 185–188, Toronto, Ontario, Canada, May 1991.
3. P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward. Towards a universal speech recognizer for multiple languages. In *Proceedings of the ASRU Workshop*, Santa Barbara, CA, December 1997. to be published.
4. J. Kohler. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In *Proceedings of the IEEE ICASSP*, volume 1, pages 417–421, Seattle, WA, May 1998.
5. D. Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, MS, June 1995. Morgan Kaufmann Publishers, Inc.
6. S. Miller, D. Stallard, R. Bobrow, and R. Schwartz. A fully statistical approach to natural language interfaces. In *Proceedings of the 34th Annual Meeting of the ACL*, pages 55–61, Santa Cruz, CA, June 1996. Morgan Kaufmann Publishers, Inc.
7. K. Papineni, S. Roukos, and T. Ward. Feature-based language understanding. In *EUROSPEECH 97*, volume 3, pages 1435–1438, Rhodes, Greece, 1997.
8. S. Della Pietra, M. Epstein, S. Roukos, and T. Ward. Fertility models for statistical natural language understanding. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 168–173, Madrid, Spain, July 1997. Morgan Kaufmann Publishers, Inc.
9. A. Ratnaparkhi, S. Roukos, and R. Todd Ward. A maximum entropy model for parsing. In *Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 803–806, Yokohama, Japan, September 1994. The Acoustical Society of Japan.
10. C. Waast, J.M. LeRoux, L. Bahl, E. Epstein, B. Lewis, P. de Souza, and S. De Gennaro. A method for modeling liaison in a speech recognition system for french. to appear in Proc. of ICSLP98, Sydney, Australia, 1998.
11. F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke. A study of multilingual speech recognition. In *EuroSpeech97*, page 359, Rhodes, Greece, September 1997. ESCA.