A Proof-of-Concept for Orthographic Named Entity Correction in Spanish Voice Queries

Julián Moreno Schneider[1(✉)], José Luis Martínez Fernández[2], and Paloma Martínez[1]

[1] Computer Science Department, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Madrid, Spain
jmschnei@inf.uc3m.es, jmartinez@daedalus.es
[2] DAEDALUS – Data, Decisions and Language S.a., Avda. de La Albufera 321, 28031 Madrid, Spain
pmf@inf.uc3m.es

Abstract. Automatic speech recognition (ASR) systems are not able to rec-ognize entities that are not present in its vocabulary. The problem considered in this paper is the misrecognition of named entities in Spanish voice queries introducing a proof-of-concept for named entity correction that provides alter-native entities to the ones incorrectly recognized or misrecognized by retrieving entities phonetically similar from a dictionary. This system is domain-depen-dent, using sports news, especially football news, regardless of the automatic speech recognition system used. The correction process exploits the query structure and its semantic information to detect where a named entity appears. The system finds the most suitable alternative entity from a dictionary previ-ously generated with the existing named entities.

Keywords: Automatic speech recognition · Audio transcription · Question answering · Phonetic representation · Named entity correction · Machine learning

1 Introduction

Automatic speech recognition (ASR) technology can be integrated in information access systems to allow searching on multimedia contents. But, in order to assure an adequate retrieval performance it is necessary to state the quality of the recognition phase, especially in speaker-independent and domain-independent environments.

Particularly important is the case in which named entities are not recognized because the information access system works with incomplete input data and is not able to find any useful information.

ASRs are not able to recognize entities that are not present in its vocabulary so the problem considered in this paper is the misrecognition of named entities in Spanish voice queries. Most works on this area try to modify the acoustic or language models of the ASR, but sometimes there is no possibility of make any change in the ASR system, e.g. if a real-time reaction is needed so there is no time to modify the acoustic model or if some predefined system (as Android or IPhone Speech Recognition) is integrated

into an application. In this case, the problem will be addressed from that point of view: **'there is no possibility of making any change in the ASR system'**.

As can be seen in the examples of Table 1, the main problem lies in the entities that are falsely recognized, i.e. the obtained entity is not the one that was said ('Woody Allen' - 'Raúl González'), or it is not even a named entity, i.e. getting a common noun when a named entity was said ('Kun Agüero' - 'una huelga').

Table 1. Examples of misrecognized named entities

Original query	Recognized query
¿Cuál fue la última película dirigida por **Woody Allen**? (What was the last film directed by **Woody Allen**?)	¿Cuál fue la última película dirigida por **Raúl González**? (What was the last film directed by **Raúl González**?)
¿En qué equipo juega **Kun Agüero**? (Which team does **Kun Agüero** play in?)	¿En qué equipo juega **una huelga**? (Which team does **una huelga** play in?)

2 Related Work

Entity correction has not been directly addressed as an independent phase after the recognition process but it has been traditionally managed within the error correction in speech recognition. There are no specific works for named entities in the state-of-the-art because in speech recognition an entity properly recognized is as important as any other word. Thus, most of the work developed in this sense tries to correct the whole transcribed query without focusing on named entities and few studies have addressed correcting only entities. The work described in [7] should be highlighted. It does not correct entities but it performs query expansion using phonetically similar words; [1] applies high-level lexical and syntactic information based on a syllables model for error correction (using also a thesaurus and domain specific dictionaries); [2] uses a rule-based system collecting error patterns and uses them to identify error in the query; [3] studies occurrence probabilities of words in dialogue; [4] detects and corrects errors using a context-based system; [5] makes a post-processing of the ASR output to correct transcription errors by offering alternatives (selected by estimating probabilities) that are not available at the output of the recognizer and [6] lets the user to select the alternatives by means of a confusion network with a large vocabulary.

Some related work can be extracted from the Spoken Term Detection Evaluation (STD) 2006 [12], but these works are related to the search of sequences of spoken words in big audio collections. They do not perform entity correction.

3 Proposal

This paper introduces a preliminary experiment for named entity correction that provides alternative entities to those incorrectly recognized or misrecognized by retrieving phonetically similar entities. This system is domain-dependent, using sports news,

specifically football news, regardless of the automatic speech recognition system used. The correction process exploits the query structure and the semantic types of phrases to detect where a named entity appears (for instance, the query "Which team does Cristiano Ronaldo play for?" has the structure "which team does FOOTBALL PLAYER play for?" where the semantic type FOOTBALL PLAYER points a named entity susceptible of being reviewed. The detection of a misrecognized named entity is done by searching it in the previously defined dictionaries (if the dictionary does not contain the named entity then it is considered to be incorrectly recognized).

The treatment needed on these entities is essentially a correction assuming that in some cases the entity will not be correctly recognized or even is not an entity (see the previous example of "Football player").

As the main difference with the related work, it can be pointed out that this proposal is a dictionary-based system that works directly over named entities instead of trying to correct each word or the whole transcription. Considering that there is no specific work on named entity correction in Spanish voice queries the objective is to perform this correction through a post processing over transcribed queries with ASR-independence (considering that it is not possible to modify nor the ASR nor its models). The domain is limited to sports news to get the named entities dictionary.

The system must find the most suitable alternative to the entities received inside the input query. To search for these alternatives a phonetic comparison between the recognized entity (by the ASR) and the entities stored in the dictionary is used and the highest scored entity is obtained (by using string comparison measurements). This functionality (together with the system's architecture) is shown in Fig. 1.

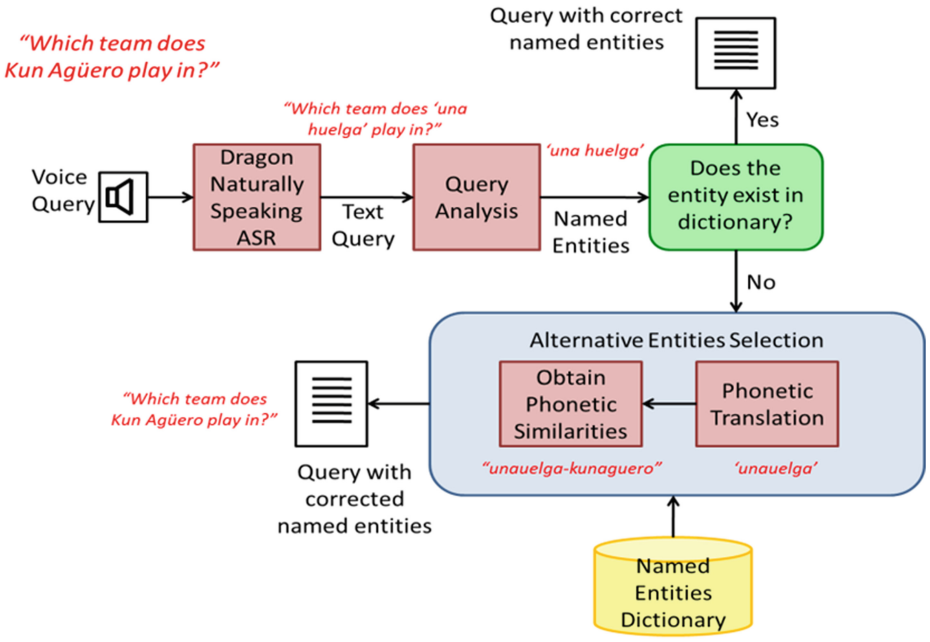

Fig. 1. Architecture of the system

The functionality of the system is structured in three main parts. Firstly, the ASR transcribes the input voice query providing a textual query. In this case two different ASR systems have been used: Dragon Naturally Speaking and Windows Speech Recognizer.

The second part of the architecture is composed by the 'entity extraction module'. I t takes care of the query analysis and searches entities inside it. This search is performed by means of a rule-based system that considers five different query patterns. These patterns are shown in Table 2.

Table 2. Available query patterns

Query patterns
¿En qué equipo juega **##JUGADOR**? (What team does **##PLAYER** play for?)
¿Quién marcó el último gol en el estadio **##ESTADIO**? (Who scored the last goal in the stadium **##STADIUM**?)
¿Quién es el máximo goleador del **##EQUIPO**? (Who was the maximum scorer of **##TEAM**?)
¿Cuántos goles ha marcado **##JUGADOR** este año? (How much goals has **##PLAYER** scored this year?)
¿Cuántos penaltis se pitaron en el último partido que se jugó en **##ESTADIO**? (How many kick goals were dictated in the last game played in **##STADIUM**?)

To determine the corresponding pattern for the input query a direct comparison is not appropriate because it can contain transcription errors. Due to that, a bag of words approach is used. It counts the number of words of each pattern contained in the input query. Once the pattern has been determined, the entity is extracted by means of its position in the query.

The next step of the system is checking whether the extracted entity has been correctly recognized or not. This functionality is performed by determining the presence of the entity in the dictionaries. If so, the entity is considered to be properly recognized and no correction is done.

Table 3. Spanish phonetic letter correspondence.

Character	Phoneme	Character	Phoneme	Character	Phoneme
a	a	k	k	t	t
b	b	l	l	u	u
c	c	m	m	v	b
d	d	n	n	w	ui
e	e	ñ	N	x	ks
f	f	o	o	y	i
g	g/j	p	p	z	z
h	–	q	k	Blank	–
i	i	r	r/R		
j	j	s	s		

On the contrary, if the entity does not appear in the dictionaries, then alternative entities for that incorrectly recognized (or misrecognized) entity are provided. A phonetic representation of the input entity is generated using a rule-based system implemented as an adaptation of the work of J. Gil [8] and LivingSpanish [11] for Spanish phonetic letter correspondence. This representation is shown in Table 3.

The similarity between the phonetic representation of the recognized entity and the phonetic representation of the named entities of the dictionary is evaluated. Indeed, several measures have been tested, such as Euclidean, Monge-Elkan, Levenshtein, Needlemann-Wunsch, Smith-Waterman, Gotoh or Smith-Waterman-Gotoh, Jaro, Jaro-Winkler and Soundex distances. A complete description of these measurements can be found in [9, 10].

The implemented dictionary is composed by a set of named entities together with its associated information (in XML format). It structure can be seen in Table 4.

Table 4. Example of named entity stored in dictionary.

```
<dictionary>
        <properties>
                <totalentities>2000</totalentities>
                <totalFP>1900</totalFP>
                <totalFS>42</totalFS>
                <totalFT>42</totalFT>
                <searchedentities>2345</searchedentities>
                <searchedFP>2200</searchedFP>
                <searchedFS>85</searchedFS>
                <searchedFT>60</searchedFT>
        </properties>
        <entities>
                <entity>
                        <text>Lionel Messi</text>
                        <type>FootballPlayer</type>
                        <popularity>0.9</popularity>
                        <historic>4</historic>
                </entity>
                ...
        </entities>
</dictionary>
```

The dictionary is composed by a set of properties and 2004 different named entities. The properties are total number of stored entities, total number of each type of entity (players, stadiums and teams) and the times that each type of entity has been selected as a suitable alternative. The entities are divided into 1874 football player names, 42 stadium names and 88 team names.

Besides, each entity is composed by its associated text, the type it belongs to, a popularity score defined by an expert and the number of times it has been selected as alternative.

4 Preliminary Evaluation

The first evaluation was carried out using 168 Spanish voice queries read by 7 different users. These queries are uniformly distributed over the five query patterns. Some examples are shown in Table 5.

Table 5. Examples of input queries read by users.

Original query	Transcribed query (with one ASR)
¿En qué equipo juega juan josé collantes?	El equipo, Juan José Collantes
¿Quién marcó el último gol en el estadio los pajaritos?	Quien marcó lo temor en el estadio los pajaritos
¿Quién es el máximo goleador del valencia?	Quién es el máximo goleador del Valencia

The queries have been transcribed using both ASRs. The first ASR was used with four different acoustic models trained with videos of different length. The first model was not trained; the second was trained only with approx. 5 min of sport news videos; the third model was trained with 50 min of sport news videos; and the fourth was trained with 40 min of football news videos. The second ASR (WSR) was not trained and only its default model was used.

The first test is performed to validate the functionality and performance of the entity classification module. In order to do that, the entities were manually extracted from the transcribed queries and then classified into types for using them as reference. The five different ASR models are tested and four different classification techniques are shown. The first technique is a direct comparison between the transcribed query and the patterns; the second is a bag-of-words technique that uses all the words (including the entity tags (#footbalplayer)); the third improves the bag-of-word technique by eliminating the tags; and the last performs a phonetic comparison between the query and the patterns.

The results obtained by the entity type classifier and the entity extractor are shown on the next table (Table 6). As can be seen, the phonetic comparison classification is the better.

Table 6. Results of entity classification module validation

	WSR	DNS - 1	DNS - 2	DNS – 3	DNS – 4
Direct comparison	0	0	0	0	0
Complete BoW	78,57 % (132)	70,83 % (119)	72,62 % (122)	73,21 % (123)	58,33 % (98)
Limited BoW	82,74 % (139)	64,88 % (109)	72,02 % (121)	70,83 % (119)	54,17 % (91)
Phonetic comparison	88,69 % (149)	86,9 % (146)	90,48 % (152)	89,88 % (151)	77,38 % (130)

The second test was performed to validate the phonetic representation system. The phonetic representation that was finally used works properly as long as the entities are 'Spanish' entities while it fails with entities from other languages. Table 7 shows some examples.

Table 7. Examples of phonetic representation of named entities

Named entity	Phonetic translation	Expected correct translation?	Correct translation
Cristiano Ronaldo	kristianoronaldo	Yes	Yes
Schweinsteiger	scuieinsteijer	No	No
HamitAltintop	amitaltintop	No	Yes
Lionel Messi	lionelmessi	Yes	Yes

The corpus used for that task was composed by 168 entities. These entities were introduced into the phonetic representation system and the output of each entity was manually revised to determine if it was correctly represented. The amount of correct represented entities was 150 entities. That leads to an accuracy of 89,29 % (Table 8).

Table 8. Phonetic representation accuracy

Phonetic representation accuracy	89,29 %

It can be remarked that all the entities that were Spanish names were correctly represented while the errors occur when there is a foreign entity (*giorgioventurin-jiorjiobenturin, ilijanajdoski-ilijanajdoski*). This is a known problem of the phonetic representation module since it has been only implemented for entities in Spanish.

The last test validates the module that retrieves alternative entities using the same corpus of 168 entities. For this purpose some different phonetic distance measurements where used. The best comparison measurements for phonetic comparison are Levenshtein Distance and Monge-Elkan-Levenshtein Distance obtaining figures near to 56.55 % in Top@10 (Levenshtein) and 50.60 % in Top@1 (Monge-Elkan-Levenshtein) (Tables 9 and 10).

Table 9. Preliminary evaluation results for entity correction using WSR

Precision without entity correction	Precision with entity correction		
		Levenshtein	Monge-Elkan-Levenshtein
30.95 %	Top@1	49.40 %	50.60 %
	Top@3	52.38 %	53.57 %
	Top@5	52.98 %	54.17 %
	Top@10	56.55 %	55.36 %

Table 10. Preliminary evaluation results for entity correction using DNS

Precision without entity correction	Precision with entity correction		
		Levenshtein	Monge-Elkan-Levenshtein
19.05 %	Top@1	38.10 %	38.10 %
	Top@3	42.86 %	41.07 %
	Top@5	42.86 %	41.07 %
	Top@10	43.45 %	42.26 %

As can be seen in the previous results, the phonetic entity correction system increases the accuracy in both cases: using WSR it increases 19,65 % and with DNS the increment is a 19,05 % for the total amount of entities.

Table 11 shows the results of the entity correction module using a system with multiple dictionaries, i.e. it uses a different dictionary for each entity type. It depends on the performance of the entity type extraction but increases the entity correction in 3 % approximately.

Table 11. Results knowing the type of the entity to be corrected

	Precision without entity correction	Precision with entity correction	
		Levenshtein	Monge-Elkan-Levenshtein
WSR	30.95 %	52.98 %	51.79 %
DNS	19.05 %	41.67 %	40.48 %

These are promising results considering that it is an ongoing work.

5 Some Conclusions

The correction of named entities in IR systems accessed by voice is absolutely necessary. There are mainly two reasons for that; on the one hand the entities are an essential unit of information for IR systems, on the other hand in most scenarios the acoustic and language model of the ASR cannot be modified to improve the results, letting this to a post processing after the recognition process.

Some comparison distance measurements were tested and finally only two of them were selected for final tests. These measures have proved very useful when making phonetic comparison. Additionally, the results are even improved when the arithmetic mean between both measures is used as a new measure (Monge-Elkan-Levenshtein).

Storing the database entities in an XML file was a good decision covering both objectives, on the one hand having a structured design, on the other hand allowing an easy understanding.

After a preliminary evaluation, the 52 % (approx.) of entity alternatives are right choices, and after making a qualitative assessment, it can be said that whenever the entity has not been recognized, the system will be able to offer an appropriate alternative.

This work has got promising results but is still in an early development stage. There are some improvements that can be outcome to the system. The first improvement would be the adaptation of the phonetic representation system to take into account different pronunciations (accents) and especially words in other languages. Besides, some ASR's return acronyms and the phonetic expansion of these acronyms could be useful for the desired purpose.

The way the system uses to determine if the entity has been correctly recognized must be further studied. Now it considers a correctly recognized entity if it is present inside the dictionaries. A phonetic comparison together with a threshold could be better.

The number of input queries that the system recognizes is now limited to 6 and it must be increased as well as the amount of entity types (now limited to three). In general, making some tests on different domains or adding more than one domain at the same time is a crucial point in order to validate the system.

Acknowledgments. This work has been partially supported by the Regional Government of Madrid under the Research Network MA2VICMR (S2009/TIC-1542) and by the Spanish Center for Industry Technological Development (CDTI, Ministry of Industry, Tourism and Trade) through the BUSCAMEDIA Project (CEN-20091026).

References

1. Jeong, M.: Using higher-level linguistic knowledge for speech recognition error correction in a spoken QA dialog. In: Proceedings of the HLT-NAACL Special Workshop on Higher-Level Linguistic Information for Speech Processing, pp. 48–55 (2004)
2. Kaki, S., Eiichiro Sumita, and Hitoshi Iida.: A Method for Correcting Speech Recognition Using the Statistical features of Character Co-occurrence, COLING-ACL'98, 653–657 (1998)
3. Ringger, E.K., Allen, J.F.: A fertility model for post correction of continuous speech recognition ICSLP'96, pp. 897–900 (1996)
4. Sarma, A., Palmer, D.: Context-based speech recognition error detection and correction. In: Proceedings of HLT-NAACL (2004)
5. Ringger, E.K., Allen, J.F.: Error correction via a post-processor for continuous speech recognition. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 427–430, Atlanta, GA (1996)
6. Ogata, J., Goto, M.: Speech repair: quick error correction just by using selection operation for speech input interfaces. In: Proceedings of Eurospeech'05, pp. 133–136 (2005)
7. Reyes-Barragán, A., Villaseñor-Pineda, L., Montes-y-Gómez, M.: Expansión fonética de la consulta para la recuperación de información en documentos hablados. Septiembre, 2011 Procesamiento del Lenguaje Natural, Revista nº 47, pp. 57–64 (2011)
8. Gil, J. Transcripción fonética: Representación escrita de los sonidos que pronunciamos. Fonética para profesores de español: De la teoría a la práctica. p. 547. Arco/Libros (2007)

9. Gusfield, D.: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York (1997)
10. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: Proceedings of II Web 2003 – IJCAI Workshop on Information Integration on the Web, pp. 73–78 (2003)
11. LivingSpanish: Correspondencia de fonemas y grafías en español. http://www.livingspanish.com/correspondencia-fonetica-grafia.htm (2011)
12. Fiscus, J.G., Ajot, J., Garofolo, J.S., Doddington, G.: Results of the 2006 spoken term detection evaluation, pp. 45–50 (2007)