Recognition and Classification of Numerical Entities in Basque

Ander Soraluze, Iñaki Alegria, Olatz Ansa, Olatz Arregi and Xabier Arregi
IXA Group. University of the Basque Country
ander.soraluze@ehu.es, i.alegria@ehu.es, olatz.ansa@ehu.es,
olatz.arregi@ehu.es, xabier.arregi@ehu.es

Abstract

This paper presents a system based on Finite State Technology that recognises and classifies numerical entities in texts written in Basque. The system deals with a wide range of entities, such as temporal expressions, numbers related to units of measurement, or those that refer to common nouns. The system obtains 86.96% F-measure score following MUC evaluation and 78.82% using IREX and CONLL simple scoring protocol.

1 Introduction

Named Entity Recognition and Classification (NERC) has become an important sub-task in the Natural Language Processing area. It is known that an effective treatment of Named Entities can benefit the performance of applications like Machine Translation (MT), Information Extraction (IE), Information Retrieval (IR) or Question Answering (QA). In the early stages, NERC systems identified a few types of entities, namely person, organisation and location names. Over time, numerical and temporal expressions have been also considered as identifiable types of entities.

Concerning to Basque, there is a NERC system called *Eihera* (Alegria et al., 2003) that recognises and classifies person, organisation and location names, but it does not deal with numerical entities up to date. The Numerical Entity Recogniser and Classifier for Basque (NuERCB) presented here aims to address this lack.

NuERCB identifies the numbers of the text and decides whether they express date or time, or are associated with units of measurement or, otherwise, just refer to common nouns. When numbers are linked to units or symbols of measurement, NuERCB determines which specific property maps with each of them. For instance, units like "square meter", "meter per second squared", "second" or "Celsius" are associated with properties like "area", "acceleration", "time" and "temperature" respectively.

Since numerical expressions, particularly those related to units of measurement, are very common in technical texts, we have used in this work the *ZT* corpus, a Basque corpus specialized in science and technology (Areta et al., 2007). In this dataset numerical expressions are more likely to appear, so it allows us to test the system on a wide variety of cases.

This paper is structured as follows. After reviewing related work, section 3 describes the linguistic features related to numerical expressions in Basque texts. Sections 4 and 5 show the methods for number detection and classification. Section 6 presents the main experimental results, which are analysed in section 7. Finally, the conclusions and future work are mentioned.

2 Related Work

The set of categories used to classify Named Entities has enriched over the time. As defined in the Message Understanding Conference (MUC) (Chinchor, 1998), Named Entity recognition consists on the identification and categorization of three types of specializations: "ENAMEX" for person, organisation and location, "TIMEX" for time and date, and "NUMEX" for money and percent. Furthermore, TIMEX2 (Ferro et al., 2003), which extends MUC definition of the TIMEX category, was used in Time Expression Recognition and Normalization evaluation (TERN 2004). Nowadays, rich hierarchies of Named Entity types have been proposed in the literature. For instance, the set of BBN[1] categories consists of 29 NE types and 64 subtypes used for Question Answering, and (Sekine and Nobata, 2004) currently gathers a hierarchy of 200 categories[2]. Temporal and numerical expressions are included in these sets.

Systems that deal with temporal and numerical expressions can be distinguished depending on the applied techniques. On the one hand, systems like LTG (Mikheev et al., 1998), MUSE (Maynard et al.,

[1] http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/
[2] http://nlp.cs.nyu.edu/ene/version7_1_0Beng.html

2001), HNERC (Farmakiotou et al., 2002), OAK (Sekine and Nobata, 2004), (Magnini et al., 2002) and (Arora et al., 2009) use pattern-based rules. On the other hand, it is worth mentioning approaches based on Hidden Markov Model (HMM) like Nymble (Bikel et al., 1997) and (Zhou and Su, 2002). A comparison between them shows that hand-crafted rule-based systems normally obtain better precision than systems based on statistical models, but the recall is lower and they require much manual work. On the contrary, statistical NERC systems require a large amount of manually annotated training data. Therefore, factors like the specificity of the domain and the availability of big training data are determinant in order to decide which method to use.

It is remarkable that systems that work on less-resourced languages normally choose a rule-based approach, as (Arora et al., 2009) for Hindi or (Farmakiotou et al., 2002) for Greek.

3 Numbers in Basque

Numbers appear in many different ways in Basque written texts. Due to Basque is an agglutinative language, a given lemma makes different word forms and this occurs even with numbers. For example, the same number can appear in different ways such as *15, 15ek, 15engana* "15, the 15, to the 15", depending on the role that it plays in the sentence.

In order to determine the different types of numerical entities we analysed the *ZT* Corpus. This corpus is a tagged collection of specialised texts in Basque. It is composed of a 1.6 million-word part, whose annotation has been revised by hand, and another automatically tagged 6 million-word part.

Numerical entities can express a wide range of information such as percentages, magnitudes, dates, times, etc. Although most of the numbers follow a simple pattern (digit and unit of measurement or category) the difficulty lies in some compound structures such as percentages or pairs of numbers with a conjunction between them. In general the patterns where the categories and the numbers are far from each other are difficult to treat. Moreover, special attention must be paid to the order of the words in the phrase. Occasionally the number can appear after the category, like in *2 lagun, lagun 2* "2 friends".

4 Number Detection

The input of NuERCB is the result of the Basque shallow syntactic analyser (Aduriz and Díaz de Ilarraza, 2003) developed in IXA[3] group. The analyser identifies and tags numbers according to six predefined types:

ZEN: Non declined numbers written with digits; cardinals *22*, percentages *% 4,5*, times *23:30*, etc.

ZEN_DEK: Declined numbers; cardinals *22k, 45i, 5ek*, percentages *% 45ean*, times *23:30etan* "at 23:30", etc.

HAUL_ZNB: Multiword numbers; *98_milioi* "98 million".

HAUL_DATA: Multiword date structure; *martxoaren_19an* "on March 19".

ERROM: Roman numerals; *VI*.

DET DZH: Numbers written in characters; *hamaika* "eleven".

We have evaluated the accuracy of the numbers detection carried out by the syntactic analyser, so that we can know the error rate in the input of NuERCB. We took 200 numbers randomly and we compared the analyser's tags with the actual ones. The obtained accuracy was 92,5%.

Observing the result, we concluded that the detection of numbers by the syntactic analyser was satisfactory as a starting point of our work. Nevertheless, some of the errors produced by the syntactic analyser have been handled by NuERCB to improve the overall performance.

5 Number Classification

In this section, we first introduce the kinds of categories used in NuERCB, and then describe the system itself.

5.1 Numerical entities

The range of categories for numerical entities is wide. On the one hand, there are categories associated with specific properties such as area, density, length, temperature, time, etc. that are represented by units or symbols: metre (m), kilogram (kg), second (s), etc. We identified 41 different properties, 2006 units and 1986 symbols. These categories are denoted as closed. On the other hand, each common noun or concept can be considered as an open category.

In the case of the closed categories, our goal is to mark numerical entities along with the property they refer to and the unit or symbol which is used for it. For example, in the sentence *Hegazkinak 2000 km/h-ko abiaduran mugi daitezke* "The airplanes can fly at 2000 km/h", 2000 is labeled with a couple of

[3] http://ixa.si.ehu.es/Ixa

tags: the symbol of measurement is "km/h" and the associated property is "speed".

In the case of the open categories, we distinguish between the percent expressions like *hazkundea % 10ekoa izan da* "the growth has been 10 %", and the simple numbers or amounts like *1250 biztanle* "1250 inhabitants". In these cases the system determines which common noun refers to the numerical entity: *% 10* is linked to *hazkundea* "the growth" and *1250* is linked to *biztanle* "inhabitants". It must be underlined that in general other systems do not classify these open categories, and in the case of percents they only tag the number followed by the percent symbol, but not the common noun that the number refers to.

5.2 System overview

NuERCB is conceived to be used in diverse applications where the response time is a critical factor. Therefore, we need NuERCB to have a high processing speed using low memory capacity. So we have chosen the Finite State Technology to implement NuERCB because of its mathematical and computational simplicity and its high performance.

NuERCB compiles a set of hand-crafted rules which have been implemented in Finite State Transducers (FST). We defined 34 FSTs to classify closed categories and 2 more for open categories that correspond to common nouns. They were defined using Foma (Hulden, 2009), an open source platform for finite-state automata and transducers. In total, the FSTs set is composed by 2095 hand-crafted rules which are able to identify 41 properties, 2006 units and 1986 symbols.

The tagging process is divided into three main phases. Firstly, the properties associated with units or symbols and boundaries of the numerical entities are tagged. Afterwards, the units or symbols of the properties that have been detected in the previous step are marked. Also, ellipsis cases of units or symbols are detected. And finally, percents, some multiword and date structures and open categories are tagged.

The input of the system is a syntactically analysed text. The format of this analysed text has been adapted to be used for the FST set, and vice versa the tagged output of the FSTs is returned to its original format.

The architecture of the system is shown in Figure 1.

To illustrate the application of the method we focus on the following examples: *21 ordu 5 minutu eta 12 segundoko* ... "... of 21 hours 5 minutes and 12 seconds" and *azalera osoaren % 8,38* "the % 8.38 of the total area". The first one is a typical composed time

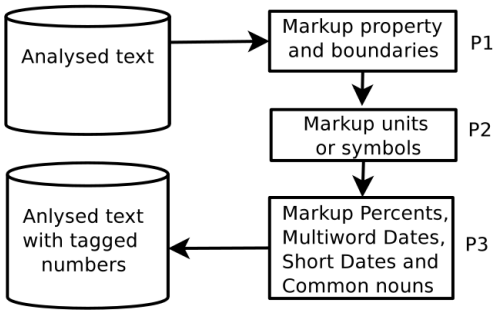

Figure 1: Architecture of NuERCB.

structure and it will be detected in the first phase (P1) and completed in the second one (P2). The second example shows a percent with a common noun category which will be detected in the third phase (P3).

P1. In this phase only the property and boundaries of the first example are detected and marked: *<TIME>21 ordu 5 minutu eta 12 segundoko</TIME>*. In Figure 2, R1[4] recognises the structure boundaries of the time property.

P2. Here units and symbols associated with each number in the structure are detected and marked: *<TIME><HOUR> 21 <MINUTE> 5 and <SECOND> 12 </TIME>*. R2 defined in Figure 2 is able to tag the *second* unit based on the *<TIME>* and *</TIME>* tags added by the previous rule (R1).

P3. Finally, the numerical entity of the second example is detected and marked. R3 in Figure 3 detects that an adjective can appear between a common noun and a percent number. Firstly the rule adds CN (Common Noun) to the tag *<PERCENT-CN> %8,38*, then a postprocess is carried out in order to replace CN by the category that corresponds. The final result is *<PERCENT-AREA> %8,38*. As we can observe the percent number is tagged correctly with the correspondent category (*azalera* "area") instead of the adjective (*osoa* "total").

6 Experimental Results

To evaluate the system we have taken 255 numerical entities and their context from the *ZT* Corpus.

The evaluation was carried out using two well known methods, the MUC evaluation system and the Exact-match evaluation which is used in IREX and CONLL.

[4]Syntax for regular expressions in Foma can be consulted in http://foma.sf.net/dokuwiki

```
define TimeStruct Number [ TimeUnit |TimeSymbol ];
define R1  TimeStruct ([(",") TimeStruct]* Conjunction TimeStruct)
                                          @-> "<TIME>" ... "</TIME>";
define SecondPost [ SecondUnit | SecondSymbol];
define R2  Number @-> "<SECOND>" ... ||"<TIME>" ?* _ SecondPost ?* "</TIME>";
```

Figure 2: Simplified rules to recognise temporal structures.

```
define R3   Number @-> "<PERCENT_CN>" ... "</PERCENT_CN>"
                                  ||CommonNoun Adjective PercentSymbol _ ;
```

Figure 3: Simplified rule to recognise percent structures.

In MUC evaluations (Grishman and Sundheim, 1996) a system is scored in two axes: its ability to find the correct type (TYPE) of the entity and its ability to find the correct text (TEXT). A correct type is credited if the entity type is assigned correctly. A correct TEXT is credited if the boundaries of the entity are marked correctly. The TYPE and TEXT are credited independently, regardless if one of them is incorrect (Nadeau and Sekine, 2007).

We use a slightly changed version of MUC evaluation. Besides the TYPE and TEXT we include SUBTYPE, which is used in the closed categories. The SUBTYPE is credited when a unit or symbol that expresses a property is marked correctly. So when we detect a numerical entity associated with a property, TYPE is credited if the property is assigned correctly, SUBTYPE is credited if the unit or symbol is marked correctly and TEXT is credited if the boundaries of the numerical entity are identified properly.

For TYPE, SUBTYPE and TEXT three measures are kept: the number of correct answers (COR), the number of actual answers that the system guesses (ACT) and the number of possible entities in the answer (POS).

In MUC, precision is calculated as COR / ACT and the recall is COR / POS. The final score is the Micro-Averaged F-measure (MAF).

IREX and CONLL share a simple scoring protocol called "Exact-Match evaluation". Systems are evaluated based on the Micro-Averaged F-measure (MAF). The precision is the percentage of named entities found by the system that are correct and the recall is the percentage of named entities present in the dataset that are found by the system. A proposed named entity is correct only if it is an exact match of the corresponding entity in the text.

In Table 1 there is a comparison between both evaluation methods taking into account two outputs. In this example, *5 metro eta 50 zentimetro* "5 metres and 50 centimetres", the MUC evaluation for the first output credits 4 points in COR and ACT: 1 point for identifying properly the structure boundaries, 1 for detecting correctly the property (*length*) and 2 more for tagging the unit of each number (*m* and *cm*). The second output is credited as follows: 3 points in COR (*length, m, cm*) and 6 points in ACT (2 boundaries, 2 times the length property and 1 point for each unit). However, Exact-Match evaluation only credits 1 point for identifying correctly all the features mentioned above in the first output.

Example	5 metres and 50 centimetres		
Correct tagging	<L> <M> 5 and <CM> 50 </L>	MUC	Exact-match
System output 1	<L> <M> 5 and <CM> 50 </L>	COR = 4 ACT = 4 POS = 4	COR = 1 ACT = 1 POS = 1
System output 2	<L> <M> 5 </L> and <L> <CM> 50 </L>	COR = 3 ACT = 6 POS = 4	COR = 0 ACT = 2 POS = 1

Table 1: Comparison of MUC and Exact-match evaluation methods.

The Precision, Recall and F-measure values obtained by NuERCB according to the two scoring protocols mentioned above are shown in Table 2. The first row shows scores for closed categories and the second one shows results for open categories. The last row summarizes the total values.

	MUC			CONLL-IREX
	P	R	F_1	F_1
CLOSED	89.59	86,95	88.25	83.70
OPEN	86.29	83.59	84.92	73.33
TOTAL	88.32	85.65	86.96	78.82

Table 2: NuERCB scores for closed and open categories.

Table 3 shows scores of the most frequent closed categories (date, time, length, weight and money), along with a specific row for percents as they have been particularly dealt among the open categories.

	MUC			CONLL-IREX
	P	R	F_1	F_1
DATE	87.18	85.00	86.08	80.00
TIME	97.73	97.73	97.73	93.10
LENGTH	90.79	92.00	91.39	92.00
WEIGHT	97.14	94.44	95.77	91.67
MONEY	92.31	88.90	90.57	88.89
PERCENT	72.73	60.38	65.98	36.84

Table 3: NuERCB scores for main closed and open categories.

The comparison of our system with other similar ones is shown in Table 4. Although systems used different category-sets, we present those that can be considered comparable.

7 Discussion

According to MUC evaluation method NuERCB obtains a 86.96% F-measure score and in conformity with Exact-Match scoring it reaches 78.82% for the total of the categories.

Analysing separately the scores for closed and open categories (see Table 2), we realize that our system's performance is better classifying closed categories (MUC: 88.25%, Exact-match: 83.70%) than open ones (MUC: 84.92%, Exact-match: 73.33%). With respect to closed categories most of the errors were due to the fact that units or symbols had not been defined in the hierarchy. As a consequence the system was not able to identify and classify these entities correctly. The problem of open categories is that sometimes the category is not near the number.

Focusing on Table 3 we notice that NuERCB gets good scores for the main categories. The lowest score are obtained in DATE and percent structure cases.

	F_1				
	1	2	3	4	5
DATE	86.08	86.98	91.9	96.59	93.73
TIME	97.73	—	92.4	92.89	87.07
WEIGHT	95.77	75.00	—	—	—
MONEY	90.57	96.47	94.83	95.54	95.47
PERCENT	65.98	—	—	94.61	98.47

Table 4: Comparison of scores among systems. 1=NuERCB, 2=OAK (Sekine and Nobata, 2004), 3=(Arora et al., 2009), 4=(Magnini et al., 2002), 5=LTG (Mikheev et al., 1998)

In DATE cases some numbers referring to date has no context clues that help in their classification. For example, the number 1963 may be a year but if there is not contextual evidence it is difficult to determine whether it is a date or not.

In the case of percent structures the task is more complex than in usual MUC systems. It is remarkable that other systems only classify simple percent structures like 20% that is a number followed by a percent symbol (%). In our case the task of identifying a percent numerical entity requires also to find the common noun that the percent number refers to. In percent structures the common noun and the percent number appear often far from each other, even in different sentences. This makes very difficult to identify correctly the category using only hand-crafted rules. Suppose that we have this example, *Emakumezkoak unibertsitateetako irakasle titularren % 13-18 soilik dira, Finlandia, Frantzia eta Espainian; Herberehetan, Alemanian eta Danimarkan % 6,5 baino gutxiago dira* "In Finland, France and Spain, women are only 13-18% of university lecturers; in Holland, Germany and Denmark are less than % 6.5." Obviously it is very complicated to tag 6.5 % with its correct category (*emakume* "woman") using just rules.

To finish the analysis of the results we compare NuERCB with other systems (see Table 4). In most of the categories our scores are similar to the others, in some cases better (TIME and WEIGHT) and in others lower (DATE and MONEY). Clearly the most significant difference is in percents as we mentioned above.

Finally, it is important to underline that some errors of the syntactic analyser, such as incorrect multiword detection or tokenizing and stemming errors, have affected our system's performance.

8 Conclusions and Future Work

We have presented the first system for Basque that addresses the recognition and classification of numerical entities. The system has a wide coverage and deals with numerical entities in a general way taking into account the diversity of phenomena in written texts. We have predefined thousands of units and symbols that allow to capture lots of properties, and we have treated common nouns as an open set of categories.

The use of Finite State Technology makes possible to process large dataset with high processing speed using low memory. We have compiled a set of 2095 hand-crafted rules in Foma. This platform facilitates the use and integration of NuERCB in information

processing applications.

Although Basque is a less-resourced language and the set of categories is not limited, evaluation scores of our system are comparable to those obtained by other systems.

In the future we aim to tackle the improvement of the performance of NuERCB in some weak points. Mainly, in what respect to percentage structures, we are considering to apply some anaphora resolution methods. In general, it will be interesting to apply machine-learning techniques like is proposed in (Erro et al., 2004) in order to correct mistakes. Using machine-learning techniques could increase the coverage of the system without rebuilding the linguistic resources.

We also aim to apply the NuERCB system in information recovery tasks, namely in an existing Question Answering system for Basque (Ansa et al., 2009). We have already integrated the NuERCB module into the QA system and nowadays we are facing its evaluation in an application-oriented way.

Acknowledgments

This work has been supported by Ander Soraluze's PhD grant from the University of the Basque Country (UPV/EHU), KNOW2 (TIN2009-14715-C04-01) and Berbatek (IE09-262) projects. Thanks to Mans Hulden for his help in defining the transducers using *foma*.

References

Aduriz, I. and Díaz de Ilarraza, A. (2003). Morphosyntactic disambiguation ands shallow parsing in Computational Processing of Basque.

Alegria, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2003). Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información, JOTRI, Madrid. 2003. ISBN 84-89315-33-7*.

Ansa, O., Arregi, X., Otegi, A., and Soraluze, A. (2009). Ihardetsi: A Basque Question Answering System at QA@CLEF 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of CLEF, 2008. LNCS, Vol. 5706/2009, pp. 369-376. Springer Berlin / Heidelberg. ISSN 0302-9743*.

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., and Sologaistoa, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. In *Copus Linguistics. Birminghan.*

Arora, S., Tyagi, R., and Arora, K. K. (2009). A Tool for Identification of Numeric, Temporal and Web Expressions in Hindi Text. In *Proceedings of ASCNT*, pages 51–57, India.

Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. (1997). Nymble: a High-Performance Learning Name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201, Morristown, NJ, USA. Association for Computational Linguistics.

Chinchor, N. A. (1998). Overview of MUC-7/MET-2. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

Erro, L. E., Solorio, T., and Computacionales, C. D. C. (2004). Improvement of Named Entity Tagging by Machine Learning. Technical report, Coordinacin de Ciencias Computacionales.

Farmakiotou, D., Karkaletsis, V., Samaritakis, G., Petasis, G., and Spyropoulos, C. D. (2002). Named entity recognition in Greek web pages. In *In Proceedings of the 2nd Panhellenic Conference on Artificial Intelligence*, pages 91–102.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2003). *TIDES 2003 Standard for the Annotation of Temporal Expressions*. MITRE corporation.

Grishman, R. and Sundheim, B. (1996). Message understanding conference - 6: A brief history. In *Proceedings of the International Conference on Computational Linguistics*.

Hulden, M. (2009). Foma: a Finite-State Compiler and Library. In *EACL (Demos)*, pages 29–32.

Magnini, B., Negri, M., Prevete, R., and Tanev, H. (2002). A WordNet-based approach to Named Entities recognition. In *COLING-02 on SEMANET*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y. (2001). Named Entity Recognition from Diverse Text Types. In *In Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark*.

Mikheev, A., Grover, C., and Moens, M. (1998). Description of The LTG System Used for MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, pages 3–26. Publisher: John Benjamins Publishing Company.

Sekine, S. and Nobata, C. (2004). Definition, dictionaries and tagger for Extended Named Entity Hierarchy. In *Conference on Language Resources and Evaluation*.

Zhou, G. and Su, J. (2002). Named Entity Recognition using an HMM-based Chunk Tagger. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.