HAREM and Klue: how to compare two tagsets for named entities annotation

Livy Real
IBM Research
livyreal@gmail.com

Alexandre Rademaker
IBM Research and FGV/EMAp
alexrad@br.ibm.com

Abstract

This paper describes an undergoing experiment to compare two tagsets for Named Entities (NE) annotation. We compared Klue 2 tagset, developed by IBM Research, with HAREM tagset, developed for tagging the Portuguese corpora used in Second HAREM competition. From this report, we expected to evaluate our methodology for comparison and to survey the problems that arise from it.

1 Introduction

Named-entity recognition (NER) is a subtask of many information extraction procedures. Its aims is to track and categorize pieces of texts (words, multiwords expressions, etc) into predefined classes such as the names of persons, organizations, etc. The state-of-the-art systems for English are able to produce near-human performance. In MUC-7 (Message Understanding Conference, 1998), the best system entering the joint evaluation scored 93.39% of F-measure while human annotators scored 97.6% and 96.95% (Perzanowski, 1998).

The good results achieved by some systems in MUC-7 don't mean that NER is entirely understood, mainly if we consider languages different from English. Moreover, to compare NER systems is a hard goal since the definition of what is a named entity itself is getting fuzzier and have passed to included not only proper nouns (Robinson, 1997). The decision to add dates, quantities or events to NE label, for example, makes necessary the retrieve of more information and is harder to keep the same score of recall and precision.

In most cases, NER is done through statistical or machine learning procedures. The IBM Statistical Information and Relation Extraction (SIRE) is one of such systems. It can be use to build trainable extractors for different domains. SIRE provides components for mention detection using Maximum Entropy models (Ratnaparkhi, 1998) that can be trained from annotated data created by using a highly optimized web-browser annotation tool, called HAT, a trainable co-reference component for grouping detected mentions in a document that correspond to the same entity, and a trainable relation extraction system.

The HAT annotation tool can be configure to use different tagset, which is also called type system, depending on the project. For news domain, a tagset named Klue was created. The Klue tagset was developed among several projects at IBM Research, mainly focused on annotate English articles with the goal extracted entities and relations between them. Therefore, Klue is a product of successive refinements, now in its third version.

After the introduction of Watson technology in the market, IBM is moving forward to make the systems adapted to work with other languages, not only English. The SIRE toolkit is part of the Watson ecosystem. Our project is to help on the improvement of SIRE models for Portuguese. Since annotated corpora were necessary to this task, we have developed an initial experiment to use an already available annotated Portuguese corpora to train a extractor model using SIRE. For this, we decide to use HAREM [1] gold collection, mapping the annotation from HAREM into Klue. Since SIRE achieves high F1 measures in many languages, this make us to believe that if we use a good annotated corpus in Portuguese, we could also obtain a good extractor using SIRE training module.

HAREM was a joint evaluation of NER system for Portuguese promoted by Linguateca, that had two editions so far. The tagset used in the gold collection of HAREM was created especially for the Second HAREM competition and it was the

[1] http://www.linguateca.pt/harem/

Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP, pages 43–49,
Beijing, China, July 26-31, 2015.

result of an agreement between the competitors that shared the combination of the types that their systems were able to recognize. In other words, the HAREM tagset was not planned as a tagset with the goal of supporting information extraction in any particular project, instead it was built from the combination of the types that several systems could annotate.

This works aims to describe our attempt to evaluate how adequate a tagset created for annotate named entities occurrences in English texts is to annotate Portuguese texts. Although Klue tagset is supposed to be a language-independent tagset, the differences we found between Klue and HAREM type systems grew some important questions: (i) Can tagsets really be language-independent? (ii) Can we believe in a true universal tagset which capture NE from any language? (iii) Does it make sense to expect that they will be completely interchangeable?

For now, we are still working on these answers and analysing both Klue and HAREM tagsets under these thoughts. For sure, tagsets are more useful if they are created to a specific domain and project, in a specific language, to a specific textual genre, but the general attempt to reach an universal tagset is an imposed challenge, since most of the tools for Natural Language Processing (NLP) aim to be universal, i.e., they aim to work with many languages and to be interoperable.

We expect to achieve a straightforward methodology to compare and adapt two different tagsets for NER. Also, we expected that there problems when using Klue tagset into Portuguese data will arise. We'll attempt to produce an empirical overview of this kind of adaptation, that is common in NLP studies, but it is not so frequently considered.

This short paper is being written while the experiment is still undergoing, but we intend to report our experience so far and share ideas with the NER researchers community.

The work is presented as following: first we'll introduce Klue and HAREM. In Section 3.1 we'll describe our proposal for the comparison between them and present the issues we found, focusing on linguistic problems from multilingual perspective, and what we could learn until now from this experiment. Finally we'll discuss some possible conclusions from it and what we leave as future work.

2 SIRE and Klue

Klue stands for Knowledge from Language Understanding and Extraction and it is one type system created to be used with SIRE toolkit. SIRE implements the Maximum Entropy Modeling for Named Entities recognition (Ratnaparkhi, 1998).

The framework that uses Klue is described in (Florian et al., 2004) and, in the last two Automatic Content Extraction (ACE) evaluations, [2] achieved top-tier results in all three evaluation languages that is participated (English, Arabic and Chinese). In ACE'02, the IBM system achieved the best values for NE detection. It achieved a F-measure of 0.685 for Arabic, 0.686 for Chinese and 0.734 for English, which is very close to human performance on this task.

Klue is used to entities tag, but also to tag relations and co-reference between them, what gives to Klue a particular feature: named entities, or *mentions*, are understood as a more open concept which includes not only proper nouns, but also pronouns, values and verbs, as it is defined "actual words referring to a certain thing or interest". This feature makes of Klue a very expressive tagset when we compare with others type systems, because it is prepared to capture much more elements than what we usually call named entity. In the research, we used the version 2 of Klue tagset, called Klue 2.

Using KLUE, *Mentions*, after a POS tagger first trial, can be categorized into *entity types*, *roles* and *sub-types*. In that sense, *entity type* indicates what type of an entity a said mention refers to, without considering context. *Entity types* have context-free nature and every *mention* with the same spelling belongs to the same *entity type*. The contextual meanings of said mention is given trough *roles*: the context sensitive meaning of a *mention*. A*subtype* is a finer-grained typological information to *entities types* which can produces *subtypes*, following the architecture TYPE.SUBTYPE, which seems to be freely inspired by Generative Lexicon strategy (Pustejovsky, 1995). Table 1 list the Klue 2 entity types and sub-types.

3 HAREM

HAREM (Cardoso, 2008a) is a joint evaluation of entities mentioned recognition systems for Por-

[2] For a description of the ACE evaluation, see `http://www.nist.gov/speech/tests/ace/`.

age	animal	award
cardinal	date	degree
disease	duration	email
event	facility	food
geologicalobj	gpe	law
location	measure	money
ordinal	organ	organization
people	person	percent
personpeople	phone	plant
product	substance	ticker
time	titlework	vehicle
weapon	weather	web

Table 1: Klue 2 Types and Subtypes

tuguese organized by Linguateca team. [3] In short, it is an initiative that aims to evaluate the success in identifying and classifying proper names in Portuguese. The set of HAREM evaluations was made between 2004 and 2007.

These evaluations came from three instances of HAREM editions: First HAREM (2004-2006), Mini HAREM (2006) and Second HAREM (2006-2008). The differences between these three evaluations are deeply explained in (Cardoso, 2008a, p.1-7). Here we use the tagset and golden collection [4] from Second HAREM, an edition made only by lusophone scholars and whose labels regarding time are more consistent than in the previous evaluations.

The Second HAREM collection includes 1,040 documents covering Brazilian and European Portuguese. Its Golden Collection is a subset of it consisting in 129 documents (2,274 paragraphs; 147,991 words) which represents 12% of the general collection. It was manually annotated and deeply discussed and revised by the HAREM team.

An important feature of HAREM, when it is compared to Klue, is the assumption that the meaning of a NE is defined only in context and can not be lexically defined. This consequently results in the fact that a NE may be marked as belonging to more than one category, especially when the context is not enough to define its meaning.

Since HAREM defines named entities as proper nouns, a very vague definition, some criteria were used for identify NE. The criteria for tagging a named entity used by HAREM includes: capitalized words (as *Obama, Lula*), expressions of time (month, dates), address, treatment pronouns (such as *Lord, Mr.*) and what they call "abstraction" (such as *illness, state, syndrome*).

HAREM categorizes named entities into *Categoria* ('category'), *Tipo* ('type'), *Subtipo* ('subtype') and also offers to annotators other possible tags, not very used on the final evaluation, as *COREL* ('co-relation') and *TIPOREL* ('type relation').

Tagging in HAREM consists in assigning at least a *category* to a named entity. After it, *types* which belong to the assigned *category* can also be assigned on the named entity, as well as *subtypes* that belongs to the same tagged *type*. We can see the HAREM annotation on the example below: [5]

```
<p>Com a influencia do <ALT>
<EM ID="hub-83689-179"
    CATEG="PESSOA" TIPO="CARGO">
bispo de Burgos</EM>
|
bispo de
<EM ID="hub-83689-180" CATEG="LOCAL"
    TIPO="HUMANO" SUBTIPO="DIVISAO">
Burgos</EM>
</ALT>
conseguiram a aprovacao do projecto
por parte de
<EM ID="hub-83689-144" CATEG="PESSOA"
    TIPO="INDIVIDUAL">Carlos V</EM>.</p>
```

where we can see entities marked with `EM`, the tag `ALT` do signal alternative annotations. In each `EM` tag we have the attributes for `CAT`, `ID` and `TIPO`. The tag `p` is the HTML tag for paragraphs.

In Table 2, we present all HAREM *categories* and *types*, we did not include the *subtypes* in this table because they are too many and not so relevant to the present work.

The systems that participated on Second HAREM are CaGE2, DobrEM, PorTexTO, Priberam, R3M, REMBRANDT, REMMA 3, SEI-Geo, SeRELeP and XIP-L2F/XEROX. Second HAREM evaluation allowed each system to choose a different task (for example, one could choose which categories to tag), what, following the authors, makes it's evaluation a bit superficial. Nevertheless the main task (to recognize a named entity and correctly classify them) is the same for all participant systems. The system with the best F measure (0.5711) was Priberam system (Amaral, 2008), followed by REMBRANDT System (Cardoso, 2008b) with its better run achieving 0.5674 F measure. All the other systems did not get a F

[3] `http://www.linguateca.pt/`

[4] The set of documents used for training the models.

[5] 'The treaty of Tordesillas divided the world.'

Category	Type
abstraction	discipline state idea name other
happening	ephemeris event organized other
person	position positiongroup indgroup membergroup individual member people other
thing	class class member object substance other
location	physical human virtual other
work	art plan reproduced other
time	duration frequency generic calendar other
value	classification currency quantity other
other	

Table 2: HAREM - Categories and Types

measure value higher than 0.5.

3.1 Comparison

To produce a comparison between Klue and HAREM tagsets, we have started from HAREM, as the annotated corpus that we want to adapt already use HAREM golden collection. Since the tagsets use different architectures, we produced a mapping table focusing in the tagset used by HAREM.

In the mapping, if a *category.entitytype* from HAREM has a straightforward relation to an *entitytype* from Klue — as the case of VALUE.QUANTITY which tags the same set of NE than the Klue *entity type* MEASURE — it is tracked. If a *category.entitytype* from HAREM has a straightforward relation to an *entitytype.role* from Klue — as the case ORGANIZATION.COMPANY and ORGANIZATION.COMMERCIAL — it is also tracked.

The complete mapping in showed in the Table 3 in the end of this report. We use **various** whenever any of the following types can be used: ANIMAL, PRODUCT, LAW, ORGANIZATION, VEHICLE, WEAPON, OTHER.

Once the mapping from Table 3 is defined, the most difficult remain task to make the translation is to collect the annotations made in-line in the HAREM documents to construct the SIRE documents format. Although both formats adopt a XML-like style, Klue docx format does not mark annotations in-line with the text, the docx document format has a special section for mentions with references to the offsets (begin and end) in the text of each mention. [6]

4 Issues

The main problem we have to deal with is how these two tagsets treats named entities. HAREM uses the more basic definition, in other words, it focus on proper names. Klue is more interested in co-reference and relations, then it is a typology that also includes common nouns, pronouns and verbs. Many pairs *category.entity* from HAREM has a straightforward *entity type* in Klue (as the relation PERSON.INDIVIDUAL into PERSON), but many other have not. ABSTRACTION.IDEA, for example, does not have a correspondent in Klue. It happens the same to the OTHER.OTHER category in HAREM, as Klue does not have so open types, there is not relation to be tracked. Whenever there is not possible relation between something tagged by HAREM into Klue, we leave the correspondence blank and the named entities marked by HAREM as belongs to these categories are not considered by our work. The elements that are not under our comparison represents 6% of the entire HAREM corpus.

In another hand, the HAREM pair *THING.OBJECT* has many possible correspondent tags in Klue (as ANIMAL, PRODUCT, LAW, ORGANIZATION, VEHICLE, WEAPON or OTHER), since the criteria used by the two tagsets are different. HAREM categorizes as THING every object or animal which is not a person and by OBJECT things with names. Klue

[6] The code that we used to translate the HAREM documents to Klue documents is available at `https://github.com/arademaker/harem`.

does not have a 'thing' category and tag directly the mention as the function of it in the real world. How to solve it is maybe the main issue that arose for our automatic methodology, as the system can not automatically choose between all these possibilities which is the correct one and we tried to avoid manual annotation in this case.

The cited issues come from the different criteria adopted by the tagsets, but also some language specific issues arose. For example, the HAREM *category* PERSON can be tagged also as several *types*: POSITION, POSITIONGROUP, INDGROUP, MEMBERGROUP, INDIVIDUAL, MEMBER, PEOPLE and OTHER. Otherwise, within Klue, there are tree different *entities types*: PERSON, PEOPLE, PERSONPEOPLE. As in English the distinction between count and non-count nouns are much more rigid and static than in Portuguese, a system prepared for English must include this distinction in its very first classification. In Portuguese, this feature is more flexible and generally defined only in the syntax level, which is not considered by *entity types* in Klue, since its classification is a context free one. It is interesting to note that this distinction in English is marked at the context free level, which is something impossible for Portuguese.

To make a relation between Klue and HAREM, many *entity types* in Klue were related to *categories*, *category.type* or *category.type.subtypes* tags in HAREM.

Although the objectives of Klue and HAREM are similar – being a tagset to be used to the classification of named entities – what is focused on each typology strategy is very different and it makes the two tagsets very distinct.

Klue has a very clear distinction between the general meaning of a *mention* (represented by *entity types*) and its contextual meaning (*role*). Within Klue, a word must have always the same *entity type* and its *role* can vary depending on the content. HAREM denies the need of having a context free meaning in NER process, since its more basic tag already depends on the context, even in cases of homophones words or expressions.

For example, 'dog' in Klue is always from the *entity type* ANIMAL and can have various *roles*: when used in a generic context, it belongs to the *role* PEOPLE as in 'Dogs are cool'; when used individually, it is tagged as PERSON, e.g. 'My dog is so cool'. Within HAREM, 'dog' in the first sentence in tagged by THING(*category*)/CLASS MEMBER (*type*); and in the last sentence 'dogs' is tagged THING(*category*)/ OBJECT (*type*).

5 Conclusions and future work

We described, in this short paper, an undergoing experiment that aims to compare two different tagsets used to NER. For now, we proposed a comparison table between them and already presented some relevant issues that we have to address before continuing the experiment.

Most issues lie in the different architectures adopted by each tagset, but specific tags which are not really language-independent, as one could expect, are also a challenge. Since Klue is not language specific (and created mainly by English speakers), it has categories which are not so relevant to Portuguese analysis.

Besides the architecture of the two chosen tagsets being different, we compared it focusing on which set of named entities each tag from HAREM included and tried to find the same set in Klue. This methodology seems to be more useful than trying to connect them finding correspondences in the architecture level. We hope that this heuristic solves both kind of problems.

What we still leave to be done is the final part of this experimente which consists in training a model in SIRE with the Golden Collection from HAREM translated to Klue tagset and evaluate the performance of SIRE comparing its results with the tools evaluated in HAREM.

References

Carlos Amaral; Helena Figueira; Afonso Mendes; Pedro Mendes; Claudia Pinto; Amaral. 2008. A workbench for developing natural language processing tools. In *Pre-proceedings of the 1st Workshop on International Proofing Tools and Language Technologies*.

Diana Santos & Nuno Cardoso. 2008a. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avalição conjunta na area*. FCCN, Portugal.

Nuno Cardoso. 2008b. Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In Cristina Mota; Diana Santos, editor, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 195–211.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N Nicolov, and S Roukos.

2004. A statistical model for multilingual entity detection and tracking. In *Proceedings of HLT-NAACL 2004*, pages 1–8.

Elaine Marsh; Dennis Perzanowski. 1998. Muc-7 evaluation of ie technology: Overview of results. *Muc 7 Proceedings*.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Adwait Ratnaparkhi. 1998. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania.

N. Chinchor; P. Robinson. 1997. Muc-7 named entity task definition. *Message Understanding Conference Proceedings*.

HAREM Category	HAREM Type	HAREM Subtype	Klue Entity Types
person	individual		person
time	calendar	date	date
location	human	discipline	gpe
organization	institution		organization
organization	administration	organization	governamental/mutigov/political
location	human	country	gpe country
person	membergroup	personpeople	
thing	class		**various**
abstraction	discipline		
value	quantity	measure	
location	human	construction	gpe facility
happening	organized		event
work	plan		law
organization	company		organization commercial
person	position		person
work	reproduced		titlework
other			
time	generic		time
happening	ephemeris		event
abstraction	name		
location	human	region	gpe
thing	object		**various**
abstraction	idea		
time	frequency		time
value	currency		money
time	duration		duration
time	timecalend	interval	date
person	people		personpeople
happening	event		event
work	reproduced	book	titlework
work	art		titlework
valor	classification		ordinal
local	physical	region	geologicalobject
time	timecalend	hour	time
person	groupind		personpeople
location	human	street	gpe
location	virtual	site	web
work	reproduced	music	titlework
organization	institution	sub	organization
work	reproduced	movie	titlework
person	groupposition		people
location	physical	watermass	geologicalobject
location	virtual	comsocial	web
work	reproduced	other	titlework
organization			organization
location	human	other	gpe
organization	administration	sub	organization governamental/mutigov/political
happening			event
location			geologicalobject/gpe/web
location	physical	island	geologicalobject
location	physical	other	geologicalobject
location	physical	relief	geologicalobject
abstraction	state		disease
thing	substance		substance
thing	class member		**various**
location	physical	watermass	geologicalobject
location	physical	planet	geologicalobject
location	other		geologicalobject/gpe/web
thing			**various**
person	member		person
abstraction			
work	art	house	titlework
location	virtual	other	web
work			titlework/law
work	art	classification	titlework
location	virtual	work	web
work	reproduced	program	titlework
work	art	other	titlework
person			person
thing	other		**various**
other	other		
location	virtual		web
work	art	painting	titlework
organization	company	sub	organization commercial
work	reproduced	theater	titlework

Table 3: Comparison - HAREM and Klue