REDES at TAC Knowledge Base Population 2016 : EDL and BeSt tracks

Yoan Gutiérrez*, Salud María Jiménez-Zafra[+], Isabel Moreno*,
M. Teresa Martín-Valdivia[+], David Tomás*, Arturo Montejo-Ráez[+],
Andrés Montoyo*, L. Alfonso Ureña-López[+], Patricio Martínez-Barco*,
Fernando Martínez-Santiago[+], Rafael Muñoz*, Eladio Blanco-López[+]
*University of Alicante / Carretera San Vicente del Raspeig, 03690, Alicante, Spain
[+]University of Jaén / Campus Las Lagunillas, 23071, Jaen, Spain
*{ygutierrez,imoreno,dtomas, montoyo,patricio,rafael}@dlsi.ua.es
[+]{sjzafra, maite,amontejo, laurena, dofer,emblanco }@ujaen.es

Abstract

This document presents a description of the Entity Linking system known as REDES, which was involved into the track Entity Discovery and Linking (EDL) of the challenge TAC Knowledge Base Population (KBP) 2016. The system developed is result of the collaboration among different research projects, in particular SAM, from which specific modules were reused and adapted to the TAC scenario; and REDES project, for which the system presented constitutes the starting point. The system proposed for this challenge provides an Entity Detection process which consists of a hybrid system that involves Stanford and NLTK NER tools. For Linking entities REDES system performs a lexical indexing and in-links counts measure. Once processed the train and test datasets it has been revealed that our proposal is able to reach promising results obtaining a precision of 81.3% in the task strong mention match and 77.0% in the task strong linked mention match.

In addition, we describe REDESb, our participation in Belief and Sentiment (BeSt) track. Our interest in this new track is given because, from our point of view, it is an excellent framework to apply REDES in a different way than EDL task. Anyway, for our first participation we face other short-term objectives. On the one hand, we want to grasp a better understanding of the issues and challenges of the task. On the other hand, we propose a model based on Bayesian programing. In spite of dealing with a very preliminary version of the proposed system, the results are promising and it encourages us to follow our approach.

1 Introduction

The goal of TAC Knowledge Base Population (KBP) workshop is to encourage people to develop and evaluate technologies for populating knowledge bases from unstructured text. For this aim, it offers 5 tracks: (i) Cold Start KBP, (ii) Validation/Ensembling, (iii) Entity Discovery and Linking, (iv) Event and (v) Belief/Sentiment. We focused on the Entity Discovery and Linking (EDL)[1] track that is related with the extraction of entity mentions from textual documents in 3 languages (Spanish, English and Chinese) and linking them to an existing Knowledge Base (KB) entry. It extends the EDL track 2015 in the following aspects: the size of the source collection has been increased from 500 documents to 90,000 documents and the individual nominal mentions have been extended to all languages and entity types.

[1] http://nlp.cs.rpi.edu/kbp/2016/taskspec.pdf

Our team, as part of REDES project [2], has participated mainly in the EDL track for English and Spanish languages. We have omitted Chinese because of our lack of experience with this language. This is the first year that we participate in this track. Although we have focused especially on the linking part of the task, we have also developed a straightforward approach for the entity discovery part. In the first step, Entity Detection, we developed a method based on two well-known Name Entity Recognition (NER) tools: Stanford NER (Finkel et al., 2005) and NER module of NLTK (Bird, 2006). In the second step, Entity Linking, we made use of the text search engine library Lucene [3] and DBpedia resources (Auer et al., 2007) and we applied different approaches in order to compute the similarity.

From our point of view EDL systems are suitable as a core part of the processing that is required for labeling expressions of beliefs. For this reason, we have participated in the BeSt track for English and Spanish: sentiment and belief detection including their source and target, where sources are named entities and targets are named entities or events or relations. This is a novel track so the comprehension of objectives and challenges is a first result of our participation. For this first participation we have carried out an initial set-up based on Bayesian programming of our system, REDESb. In the long term, our interest in BeSt is understood as a way to evaluate, in future, the performance of REDES in a rather different application than EDL, such as belief labeling.

The rest of the paper is structured as follows. Section 2 shows a review of the state of the art and a description of the task, as well as its challenges. The system is described in Section 3. Following, in Section 4, a consistent description of the evaluation and results is exposed. As part of this section a discussion is presented followed by an explanation, in Section 5, of how the work presented here takes part in two different projects, i.e. collaboration between a Spanish one and an European one. Section 6 depicts our participation in Belief and Sentiment Evaluation track. Finally, conclusions and future works are outlined in Section 7.

[2] https://gplsi.dlsi.ua.es/redes/
[3] https://lucene.apache.org/

2 Related works and task challenges

The task of entity linking has attracted a lot of attention in terms of shared tasks (Cano et al., 2014; Cano, Preotiuc-Pietro, Radovanovic, Weller, & Dadzie, 2016; Ji, Nothman, & Hachey, 2014, 2015; Rizzo, Cano Basave, Pereira, & Varga, 2015).

During the past Text Analysis Conference (TAC) workshop, the Tri-Lingual EDL track (Ji et al., 2015) was proposed. Next, we review some of the most relevant systems that participated in this challenge.

The IBM team (Sil, Dinu, & Florian, 2015) presented an EDL system for the three languages. For Spanish and English, the entity discovery is based on a combination of deep neural networks (NN) and Conditional Random Fields (CRF); whereas for Chinese this step combines CRF models. The entity linking system, which was applied to all languages, is based on a maximum-entropy model. It uses language independent features, such as exact match, acronyms or Wikipedia categories, among others.

The KELVIN system (Finin et al., 2015) was extended to be multilingual. It applied the Bing translation service to translate Spanish and Chinese to English. Their linking approach compares an entity type and mentions to the external KB entity types, names and aliases. The candidate set is produced by retrieving entities whose names or aliases match to the KB. The candidates are ranked by the most used mention and a significance score.

The BUPT team (Tan, Zheng, Li, & Wang, 2015) presented an EDL system for these three languages. The discovery phase uses an existing NER (Stanford NER) to detected mentions. In addition, mentions are expanded with aliases, acronyms, etc. Their linking phase produces candidates (entries in the KB) using an Elastic Search index of the KB (Freebase). These candidates are ranked based on topic-sensitive random walk with restart.

Fauceglia, Lin, Ma, & Hovy (2015) proposed the CMU system, a unified graph-based approach based on Freebase KB for the three languages. For Spanish and English, named entities are detected as n-grams, with at least one name. These are compared against a name map indexed in Lucene. But Chinese works at character level. Then, concept disambiguation entity and linking is

performed simultaneously. This graph approach is based on semantic signatures built using Personalized PageRank algorithm with node-dependent restart.

The RPI_BLENDER team (Hong et al., 2015) used a different entity discovery approach for each language. For English, they apply regular expressions together with a linear-chain CRFs model, whereas an existing NER (Stanford NER) is employed for Spanish and Chinese. On top of that, the type of each mention is stablished using a mapping between Abstract Meaning Representation corpus and DBpedia type. Finally, their linking approach is graph-based and takes into account co-occurrence mentions within a paragraph and a surface dictionary.

After analysing these EDL systems, the following conclusions can be drawn: (i) CRF is one of the most common algorithms for entity discovery (Hong et al., 2015; Sil et al., 2015; Tan et al., 2015); (ii) there is a trend of re-using existing NER tools such as Stanford NER (Finin et al., 2015; Hong et al., 2015) ; and (iii) graph-based approaches are usual for entity linking (Fauceglia et al., 2015; Hong et al., 2015).

2.1 Task description

Entity linking is the task of matching a textual entity mention to a KB, such as a Wikipedia page, that is a canonical entry for that entity (Rao et al., 2013). For instance, given a mention in a text to "*Al Pacino*", the goal of this task is to determine that it refers to the entity described in this specific entry in Wikipedia, see this example: "*Al_Pacino*" [4]. This task is more challenging than traditional Named Entity Recognition (NER), where the goal is to determine the occurrences of names in text and their classification. In the previous example, a NER system would determine that "*Al Pacino*" is a person or that "*Los Angeles*" is a location (Nadeau and Sekine, 2007). Entity linking requires a NER system, but this process must be complemented by a following disambiguation phase where this person or location is linked to an unambiguous entity stored in a knowledge base.

Entity linking systems must face three main challenges (Drezde et al., 2010): (i) name variations, (ii) entity ambiguity, and (iii) absence.

The first one, name variations, refers to the fact that an entity often has multiple forms, including shortened forms (*Leonardo DiCaprio* / *Leo DiCaprio*), aliases (*Dwayne Johnson* / *The Rock*), alternate spellings (*Osama* / *Ussamah*) and abbreviations (*British Broadcasting Corporation* / *BBC*). Approaches to entity linking must provide a means to deal with this problem in order to achieve a good recall, i.e., retrieving a large fraction of relevant instances. A common strategy to solve this problem is the use of string similarity metrics, which measures distance between two text strings for approximate (fuzzy) string matching. Nevertheless, fuzzy matching techniques are not good to deal with aliases, alternate spelling and abbreviations. More sophisticated approaches, such as Latent Semantic Analysis (LSA), are required to this end. LSA is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Deerwester et al., 1990). The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear determines the similarity of meaning of words. For instance, using LSA a system could detect that *The Rock* and *Dwayne Johnson* refer to the same entity because these names appear usually surrounded by the same words (e.g. the names of his films). In the case of links to Wikipedia, name variations can be overcome by the use of disambiguation and redirect pages. Disambiguation pages on Wikipedia are used as a process of resolving conflicts in article titles that occur when a single term can be associated with more than one topic.[5] Redirects are usually created because readers may search for an article under different names. Examples are: alternative names for the same thing, alternative spellings, capitalizations, and common misspellings.[6] These pages can be used as cues for solving aliases, alternative spellings and abbreviations.

The second challenge is entity ambiguity: a single mention to an entity can match multiple knowledge base entries, since many entities tend to be polysemous. For instance, *Francis Bacon* can refer both to the English philosopher and to the Irish artist. Approaches to disambiguating entities

[4] http://es.wikipedia.org/wiki/Al_Pacino

[5] https://en.wikipedia.org/wiki/Help:Disambiguation

[6] https://en.wikipedia.org/wiki/Help:Redirect

do it in much the same ways as a human does: when a potentially ambiguous entity is encountered, the surrounding text is examined for contextual cues, i.e., hints that help to disambiguate an entity. In the case of Wikipedia, each page has a set of categories assigned. These categories can be used to link articles under a common topic. In the example above, the philosopher *Francis Bacon* is associated to categories "*English philosophers*", "*English essayists*" and "*Empiricists*" among others, whereas the artist *Francis Bacon* is assigned to "*Anglo-Irish artists*", "*Modern painters*" and "*Painters from London*". All this information can help in the disambiguation process.

Finally, the third challenge is to solve the problem of absence: identifying whether an entity has or not a related entry in the target knowledge base. This problem affects the precision of the system, i.e., the fraction of retrieved instances that are relevant. A system must avoid detecting entities that are not present in text. Confidence thresholds are usually employed to discard these unwanted false positives.

3 System Description

REDES system is constituted by two modules, one for detecting and classifying Named Entities (NE) and other for linking. The overall process begins when a text is introduced into the system. At this time the Entity Discovery stage acts by considering different strategies (see a detailed description in section 3.1) for detecting NEs appearing in the text, which will also be classified in categories. Having that, the Entity Linking stage (see a detailed description in section 3.2) applies different Word Sense Disambiguation strategies to determine for each NE which semantic Entry (i.e. DBpedia Entries) is more appropriate, among a large list of candidates. Figure 1 provides general architecture of the system REDES.

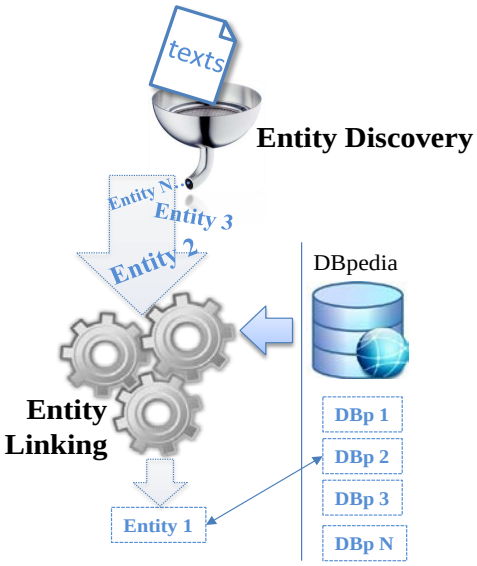

Figure 1. Overall workflow

3.1 Entity Discovery

Entity Discovery consists of two stages. In the first place, entities present in the text are detected and, in the second place, the types of the entities, which have been extracted in the previous step, are stablished. We have used two different alternatives but, previously, a preprocessing step, common for both of them, has been applied.

The system takes as input a document and parses it using the *ElementTree* library[7]. Once the text is obtained, it is passed through a NLTK language detector based on stop words. This detector counts the occurrences for each language and assumes that the language with more matches is the one the text belongs to. Once the language is detected, the entities are extracted. For this, we applied Stanford NER and NLTK NER tools following two different approaches.

Stanford NER provides a general implementation of linear chain Conditional Random Field sequence models (Finkel et al., 2005). For English texts, we used this tool with a specific model trained on the corpus *conll2003*[8] (Sang et al., 2003). Similarly, for Spanish texts, a different model trained on the AnCora corpus[9] (Taulé et al., 2008) was employed. Both models detect 4 classes: Miscellaneous (MISC), Location

[7] https://docs.python.org/2/library/xml.etree.elementtree.html

[8] http://www.cnts.ua.ac.be/conll2003/ner/

[9] http://clic.ub.edu/corpus/es/ancora

(LOC), Organization (ORG) and Person (PER). MISC is not a valid tag in the output format for EDL track. Therefore, after analyzing the data of previous years, we changed this label for the GPE tag due to the high probability of occurrence of geo-political entities.

On the other hand, NLTK NER uses a Maximum Entropy classifier trained on the ACE corpus [10] (Mitchell et al., 2005). This model detects 6 types of entities: facility (FAC), geo-political (GPE), geo-social-political (GSP), location (LOC), organization (ORG) and person (PER). GSP type is not a valid output in the task, so we assume that it refers to GPE type. This model is provided for English language. Consequently, in order to use it for Spanish language, TextBlob [11] library was employed for the translation of Spanish texts into English.

As we have mentioned before, we used these tools following two alternatives:

1. **DISCOVERY REDES: hybrid system.** NER Stanford tool is used in order to detect entities. For each entity detected, if NLTK NER tool also detects it, the type identified with this tool is returned, if not the type identified by Stanford tool is assigned to the entity.
2. **DISCOVERY REDES: NLTK system.** NLTK NER is applied in both steps, detection of the entities and identification of the types.

Once the entities have been detected, the system applies these improvements:

- *Compound words:* if there are two consecutive entities labeled, they are taken as a single entity rather than taking them as independent entities.
- *Unification:* if an entity that has been detected several times has different labels, the tags of this entity are modified by the one that appears most frequently.

According to the specifications of the EDL 2016 task, it is also necessary to distinguish between named mentions (NAM) or nominal mentions (NOM). A named mention refers to an entity by its proper name, acronym, nickname, alias, abbreviation, or other alternate name (e.g. *"Elisabeth"*). On the other hand, nominal mentions are common nouns or noun phrases that refer to an entity but which are not actually names (e.g. *"The girl wearing red trousers"*). This aspect has not been studied and the system classifies all the entities as named mentions, which is the most frequent mention type.

3.2 Entity Linking

The Linking process performed in this module counts with three alternatives. However, all of them make use of common knowledge extracted from the EDL Knowledge Base. To that end, a lexical index was created by indexing the DBpedia's URIs, the title and the description of the KB entries using Lucene library. Having that, these alternatives are able to retrieve from a Named Entity a list of DBpedia Entry candidates based on lexical similarities. The lexical similarity applied in all cases is based on the well-known Levenshtein [12] (LEV) edit distance. All linking alternatives begin having the Named Entity as text input and its context (text where it was extracted) in order to assign the right entry candidate.

We tested the following alternatives for the linking process:

1. **Linking REDES Title + Inlinks**. This alternative is the most complex one, since this combines the in-links information retrieved by directly querying DBpedia for its Entry; and the best similarity between the Entity and its Linking Candidate's titles, using our lexical index, to set a ranking of entity's Linking Candidates. Note that a Linking Candidate (which contains a title and also others attributes) is a unique Entry of DBpedia which can be associated to the different Entities present in a text. The formula which combines both measures for scoring the candidates is as follows:

$$Scoring(NE) = Inlinks(EC_i^{NE}) + (1 - LEV(NE, LinkingTitle)) \quad (1)$$

Where:

[10] https://catalog.ldc.upenn.edu/LDC2005T09
[11] https://textblob.readthedocs.io/en/dev/
[12] http://www.miislita.com/searchito/levenshtein-edit-distance.html

- NE is a named Entity,
- $EC_i \in$ Linking Candidates of NE,
- LEV represents the Levenshtein distance between two texts.

The *Inlinks* represent a normalized value (ranging from 0-1) of the in-links count of each linking candidate. For computing that, it is taken as range top the maximum count of *Inlinks* among the candidates of each Entity and it is computed the following formula:

$$Inlinks(EC_i^{NE}) = Inlinks(EC_i) / Max(NE) \quad (2)$$

Where :

- Inlinks(EC_i) obtains the number of DBpedia inlinks on a Linking Candidate;,
- Max(NE) gets the maximum number of inlinks among of Inlinks Candidates of an Named Entity (NE).

As can be seen, Levenshtein constitutes an edition distance so, while more similar the texts are the result value tends to zero. So that, aiming at scoring the candidates and selecting the most relevant one, we normalized both Levenshtein and Inlinks, in the range *0-1* before scoring these values. The Linking Candidate that computes the maximum score is selected for being linked to the Entity analyzed.

2. **Linking REDESTitles**. This alternative computes the similarity between an Entity and its Linking Candidate's titles, for scoring the best Linking Candidate taking advantage of our lexical index.

$$Scoring\ (NE) = (1- LEV\ (NE, LinkingTitle) \quad (3)$$

3. **Linking REDESTitles + Description**. This alternative computes the scoring confidence as the lexical edit distance between an Entity (considering the context where it appears, as *NEcontext*) and its Linking Candidate's titles (*LinkingTitle*) and descriptions (*LinkingDesc*).

$$Scoring\ (NE) = (1 - LEV\ (NE, LinkingTitle)) + (1-LEV\ (NEcontext, LinkingDesc)) \quad (4)$$

4 Evaluation and Discussion

We were able to submit 4 alternatives (REDES1, REDES2, REDES3, and REDES4) for the EDL task. Next subsections present these results for Entity Discovery alone, as well as Entity Discovery and Linking.

4.1 Entity Discovery

For Entity Discovery, we only applied two different approaches: hybrid system (REDES1, REDES2) and NLTK system (REDES3, REDES4). Table 1 lists our official results for our best run in terms of Precision, Recall and F1-score for the Entity Discovery stage. *NER* and *NERC* metrics evaluate mention detection and classification, respectively.

	NER			***NERC***		
Lang	**P**	**R**	**F1**	**P**	**R**	**F1**
English	**0.813**	0.417	0.552	0.621	0.319	0.421
Spanish	**0.622**	0.344	0.443	0.343	0.190	0.244
All	0.728	0.249	0.372	0.496	0.170	0.253

Table 1. Entity Discovery performance on the evaluation set in the task "*strong mention match task*" for our best run (hybrid system: REDES1, REDES2)

4.2 Entity Discovery and Linking

Once decided the Entity Discovery approach to use, we set-up our final system's configuration. Table 2 provides the combination of Entity Discovery and Linking stages that finally where involved into the EDL competition.

Approach	**Detection approach**	**Linking approach**
REDES1	hybrid (Stanford and NLTK NER)	Titles + Inlinks
REDES2	hybrid (Stanford and NLTK NER)	Titles
REDES3	NLTK NER	Titles + Description
REDES4	NLTK NER	Titles

Table 2. System's configurations

Table 3 provides the official results of REDES regarding our entity linking strategy (i.e. strong linked mention match) for the four different runs. The best Precision, Recall and F1 are obtained for English: REDES1 and REDES2 obtained the best Precision (0.77%) and F1 (0.601%); whereas REDES 3 got the best Recall (0.498%).

Approach	Lang	Linking		
		P	R	F1
REDES1	All	0.691	0.288	0.407
REDES2	All	0.691	0.288	0.407
REDES3	All	0.671	0.241	0.354
REDES4	All	0.687	0.234	0.349
REDES1	English	**0.770**	0.494	**0.601**
REDES2	English	**0.770**	0.494	**0.601**
REDES3	English	0.677	**0.498**	0.574
REDES4	English	0.690	0.490	0.573
REDES1	Spanish	0.587	0.400	0.476
REDES2	Spanish	0.587	0.400	0.476
REDES3	Spanish	0.653	0.212	0.321
REDES4	Spanish	0.676	0.197	0.305

Table 3. Entity linking results (strong linked mention match) for all four alternatives and languages tackled (best results bold-faced)

4.3 Discussion

If we analyze the whole Entity Discovery process the system is accurate in the detection of entities (see precision in Table 1), especially for English language. If we compared our results in terms of precision with those obtained last year, we would be in 5th and 6th position for English and Spanish respectively. However, the recall is very low. The system detects less than 50% of the entities. On the other hand, in relation to the classification of the entities, the precision and recall of the system are lower. Analyzing the results we found that most of the errors are related to the identification of FAC and LOC entities. Last year, FAC performance was the lowest by far (Ji et al., 2015).

Considering that our approaches for Entity Discovery and Linking take into account noncomplex algorithms (i.e. only requires a knowledge base supported by indexing techniques and existing NER tools retrained), obtaining more than 70% of precision when detecting and linking Entities suppose a promising base for EDL systems. Note that, the best Spanish precision belongs to REDES3 and REDES4 (see Table 5), but their recall is too low. So, it is better to focus on REDES1 and REDES2 results in general.

It seems the systems REDES1 and REDES2 perform similar, being the in-links information not useful for this campaign. This idea has been reused from other approaches (see project collaboration section 6) like in (Tomás et al., 2015) where they were able to obtain promising results using DBpedia in-links.

We believe that one of the issues arisen in this challenge that could affect the in-links capture was the KB indexing time. Motivated by a limited time for indexing DBpedia information (URL, title, description and in-links), we discarded the last field for the index process. In-links information was obtained from a DBpedia query (relative to the Name Entities to link) limited to 10 entries in order to obtain the Linking Candidates. Once obtained those Linking Candidates, we computed the Linking formula described in section 3.1 related to the in-links parameters. But, it could be very possible that too many candidates could be out of this scope. So that maybe, it could be the reason from which in-links information does not added valuable information for this EDL challenge.

5 Project collaboration

A part of the technologies employed in the REDES[13] system were developed in the framework of the EU-funded project SAM[14]. The goal of this project was to build an advanced digital media delivery platform, combining second screen and content syndication technologies in the domain of Social TV, where television and social media are united to promote communication and social interaction related to a broadcasted program content. In this platform an approach to entity linking on two different KBs was created. First, the system identified and linked mentions in text to related Wikipedia pages. Secondly, it also identified references to instances contained in its own media assets KB (e.g. books, songs, films, actors, etc.).

The potential customers of SAM are both business stakeholders (such as media broadcasters, content asset providers, software companies and digital marketing agencies) and end users. For the former, entity linking provides a number of benefits, including the enrichment of their contents by linking them to additional internal (media assets from the SAM knowledge base) and external (Wikipedia) sources of information. Regarding the benefits for end users, entity linking provides an augmented experience in which they can discover new information about an asset, creating richer experiences around the original contents. For

[13] http://gplsi.dlsi.ua.es/redes/
[14] http://www.socialisingaroundmedia.com/

instance, a user is watching the film *Casino Royale* in this platform, and thanks to the entity linking module, would get additional information related to actors *Daniel Craig* and *Mads Mikkelsen* from Wikipedia, and also to other related assets in the SAM platform based on the linking to its own KB, such as books created by *Ian Fleming*, the writer of the series of spy novels.

So it can be said that, in this EDL challenge, SAM technologies have played a key role to perform REDES linking approaches. From SAM some elements have been reused and adapted. For example, SAM has a semantic module which automatically indexes, with Lucene, SAM KB entries (i.e. SAM Assets). Thus, this module has been the kernel of the Linking stage of REDES. It has been adapted to the KB provided by the TAC organizers. Besides, REDES has taken advantage from SAM to automatically provide in-links scorings that allows to link a DBpedia Entry.

6 Belief and Sentiment Evaluation (BeSt)

In this section we describe REDESB, our participation in BeSt, a novel track that is intended to evaluate sentiment and belief detection with source and target, where sources are named entities and targets are named entities or events or relations.

The input of a BeSt system is a set of documents such as newswire and discussion forums where entities, mentions to these entities, relations and events are previously labelled. It is our hypothesis that an EDL system such as REDES will largely improve belief detection and identification intended as a whole system, but for this participation we have addressed our efforts to a more short-term objective: the development of a framework based on Bayesian programming from scratch. This framework is thought to enable the integration of both REDES and REDESB in only one system in a near future.

6.1 System Description

For our first participation in BeSt we have focused on belief classification rather than sentiment identification and polarity classification. Thus, for a given relation or event, it is required (i) to identify the source of the belief, i.e. who has a mental attitude towards what? and (ii) which grade of belief happens: the source is convinced regarding his or her belief (committed belief, CB), it is perceived as possible but not for sure (non-committed belief, NCB), it is not an own belief but just a reported belief (RCB) or that does not represent a belief (not applicable, NA).

For this first version of our system, we have followed a naïve approach for the identification of the source of the belief: the source of the belief is the same that the source of the document or post in which such belief happens. Consequently, if the source of the document is unknown or anonymous, then every belief will be equally unknown or anonymous.

In order to classify every belief we have developed a system following a number of assumptions as design guides:

- It is obvious that it is not possible to encode the same knowledge used by the human annotators. Thus, information to automatically classify every belief is less than the necessary.
- The number of examples for training is rather scarce: 209 posts from discussion forums and 37 documents from newswire.
- The issues that should be considered are very heterogeneous, with different levels of abstraction: the type of relation or event, the type of entities as arguments, the author of the whole text where the belief is found, the verb that the source uses to mention the belief and so on.
- The system should make easy the addition of new issues as they become available.

As a consequence of these assumptions we have implemented a system using Bayesian programming language (Jayce, 2003; Gelman et al., 2014). More concisely we have used ProBT (Bessiere et al., 2013). It is a formalism, a methodology, an API and an inference engine to solve problems with incomplete and uncertain information. This formalism requires defining a conjunction or set of variables, the sample space. The searched variable B is the category of the given belief. Then, a number of known variables $K_1,\ldots,K_n$ are given in order to estimate the probability distribution of the searched variable. For this very preliminary version of our system we consider a number of variables such as the

event/relation sub-type (K_1), the type of the entities (K_2), the arguments of these relations and events (K_3) and the POS of the trigger word related to the belief (K_4). We are interested in the probability of the conjunction of these variables:

$$P(B \wedge K_1 \wedge K_2 \wedge K_3 \wedge K_4)$$

The computation of this joint distribution is not possible if we do not accomplish certain grade of decomposition to obtain a good model, easy to compute and easy to identify. Decomposition restates the joint distribution as a product of simpler distributions. Starting from the joint distribution and applying recursively the conjunction rule we obtain:

$$P(B \wedge K_1 \wedge K_2 \wedge K_3 \wedge K_4) = P(B) \times P(K_1|B) \times P(K_2|B \wedge K_1) \times P(K_4|B \wedge K_1 \wedge K_2) \times P(K_4|B \wedge K_1 \wedge K_2 \wedge K_3)$$

We simplify it drastically by assuming that every K variable is independent from the rest of them. For example, the type of the relation (K_1) does not keep relation with the type of the argument (K_2). Then, it is possible to rewrite the previous expression as:

$$P(B \wedge K_1 \wedge K_2 \wedge K3 \wedge K_4) = P(B) \times P(K_1|B) \times P(K_2|B) \times P(K_3|B) \times P(K_4|B)$$

To be able to compute the joint distribution, we must now specify the 4+1 distributions appearing in the decomposition:
$P(B)$ is the distribution a priori of B. We have estimated this distribution by using the testbed provided by the organization:

$$P(B_i) = n_i/N,\ B_i,\ for\ each\ B_i \in \{cb, ncb, rob, na\}$$

where n_i is the number of beliefs of type B_i and N, the total number of beliefs.

$P(K_j|B)$ is the likelihood, the conditional distribution of the variable K_j for a given belief B_i. For example, the likelihood for the event/relation sub-type (K_1) given a committed belief is:

$$P(K_1|B_1) = \{P(K_1=V_1|B_1) \ldots P(K_1=V_m|B_1)\}$$

where $V_1 \ldots V_m$ are the m sub-types of events and relations and it is computed as a Laplace succession[15]:

$$P(K_1=V_j|B_1) = (1+n_j)/(M+N_1),\ 1 \leq j \leq m$$

where nj is the number of beliefs of type B_1 that presents a relation/event of type V_j, and N_1 is the total of B_1 beliefs.

Thus, the joint distribution is fully specified and it is possible to ask questions such as:

$$P(B|K_1 \wedge K_2 \wedge K_3 \wedge K_4) = P(B) \times P(K_1|B) \times P(K_2|B) \times P(K_3|B) \times P(K_4|B) / \sum_B P(K_1|B) \times P(K_2|B) \times P(K_3|B) \times P(K_4|B)$$

6.2 Results

We have applied our system to Spanish and English. Such as Table 4 shows, there is space to improve our system, since the variables that we have considered are very straightforward. We have obtained a better result for newswire documents. It could be explained because we have applied a naïve method to identify the source and newswires are anonymous and the most frequent source for the beliefs within these documents is equally anonymous. As a result, the identification of the source has a minor impact with respect of the discussion forums dataset, where the sources of beliefs are more heterogeneous.

	Discussion forums			***Newswire***		
Lang	**P**	**R**	**F1**	**P**	**R**	**F1**
English	**0.492**	0.569	0.530	0.752	0.441	0.556
Spanish	**0.440**	0.535	0.489	0.603	0.490	0.546

Table 4 Obtained results for English and Spanish. The evaluation distinguishes between discussion forums and newswire documents

[15] We use a Laplace succession to estimate the likelihoods because if we use just proportions of beliefs where a given subcategory appears, then we are assuming that the probability for that what has no yet been observed is impossible! This is the case of 'investorshareholder' that it does not appear as part of the training set but, indeed, it appears in the evaluation set.

7 Conclusion and future works

The work presented in this paper has been the result of the collaboration among different research projects from which a system for detecting and linking entities started up. The system carried out constitutes a starting point by setting a baseline for the project REDES, since basic statements for EDL have been considered. This baseline is able to reach promising results, obtaining a precision of 81.3% in the task strong mention match and 77.0% in the task strong linked mention match. These results constitute a good base for continuing improving the system since we plan to add new semantic strategies in the future.

Due to the test's results arisen from REDES1 and REDES2 do not revealed any significant improvement, we consider it is necessary to study a broader scope for capturing candidates from the DBpedia's in-links. This is very valuable information, contrasted by several background researches and needs to be studied.

Another future plan for REDES is to enrich a bit more the information indexed in the Linking Module. This enrichment would consider semantic fields collected from DBpedia like *subject* and *type*. In Addition, alternative semantic sources like Freebase[16] and BabelNet[17] will be studied, since they would add complementary knowledge to the REDES indexing bucket.

For the BeSt track, we have developed a framework based on Bayesian programming, REDESb. The Bayesian model has to be improved and the first way is the inclusion of variables representing more semantic issues. More concisely we will focus on the study of those verbs that are used by the source to express every given belief. It is necessary the inclusion of a semantic representation of the text as part of our model, such as dependency trees, and the definition of a set of rules or operands to discover which verb is the one used by the source. The next step will be the integration of REDES and REDESb where REDES will be a core part of the preprocessing of the texts previous to the identification and labelling of beliefs.

[16] www.freebase.com
[17] babelnet.org

Acknowledgments

This work has been partially supported by the Spanish "Ministerio de Educación, Cultura y Deporte" (MECD - scholarship FPU014/00983), the Spanish "Ministerio de Economía y Competitividad (MINECO)" (projects TIN2015-65136-C2-1-R / TIN2015-65136-C2-2-R), the European Commission (SAM project FP7-611312), and the Generalitat Valenciana Government (PROMETEOII/2014/001). In addition, acknowledging the support of technical staff such as Ulises Serrano, Francisco Agulló and Javier Fernández from the University of Alicante.

References

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K. & Harshman, R.. 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci., 41: 391–407.

Dredze, M. ; McNamee, P.; Rao, D. ; Gerber, A. & Finin, T. 2010. Entity disambiguation for knowledge base population. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 277-285.

Nadeau D.; & Sekine. S. 2007. A survey of named entity recognition and classification. Journal Lingvisticae Investigationes, 30, 1, 3-26.

Rao, D. ; McNamee, P. & Dredze. M. 2013. Entity Linking: Finding Extracted Entities in a Knowledge Base. In Multi-source, Multilingual Information Extraction and Summarization. Springer, 93-115.

Cano, A. E., Preotiuc-Pietro, D., Radovanovic, D., Weller, K., & Dadzie, A. S. 2016. #Microposts2016 – 6th Workshop on "Making Sense of Microposts." In WWW'16 Companion Proceedings of the 25th International Conference Companion on World Wide Web (pp. 1551–1552). http://doi.org/10.1145/2872518.2893528

Cano, A. E., Rizzo, G., Varga, A., Rowe, M., Stankovic, M., & Dadzie, A. S. 2014. Making sense of microposts (#Microposts2014) named entity extraction & linking challenge. In Proceedings of the 4th Workshop on Making Sense of Microposts (#Microposts2014) at the 23th International Conference on the World Wide Web (WWW'14) (Vol. 1141, pp. 54–60).

Fauceglia, N., Lin, Y., Ma, X., & Hovy, E. 2015. CMU System for Entity Discovery and Linking at TAC-KBP 2015, (January).

Finin, T., Mayfield, J., Gao, N., Lawrie, D., Oard, D., Lin, Y., … Dowd, T. 2015. HLTCOE Participation in TAC

KBP 2015 : Cold Start and TEDL.

Hong, Y., Lu, D., Yu, D., Pan, X., Wang, X., Chen, Y., … Ji, H. 2015. RPI BLENDER TAC-KBP2015 System Description. Proceedings of Text Analysis Conference 2015.

Ji, H., Nothman, J., & Hachey, B. 2014. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In TAC (Text Analysis Conference) 2014.

Ji, H., Nothman, J., & Hachey, B. 2015. Overview of TAC-KBP2015 Entity Discovery and Linking Tasks. In Proceedings of Text Analysis Conference 2015.

Rizzo, G., Cano Basave, A. E., Pereira, B., & Varga, A. 2015. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition & Linking Challenge. Proceedings of the 5th Workshop on Making Sense of Microposts (#Microposts2015) at the 24th International Conference on the World Wide Web (WWW'15), 1395, 44–53.

Sil, A., Dinu, G., & Florian, R. 2015. The IBM Systems for Trilingual Entity Discovery and Linking at TAC 2015.

Tan, Y., Zheng, D., Li, M., & Wang, X. 2015. BUPTTeam Participation at TAC 2015 Knowledge Base Population. In Proceedings of Text Analysis Conference 2015.

Tomás, D., Gutiérrez, Y. & Agulló, F. 2015 Entity Linking in Media Content and User Comments: Connecting Data to Wikipedia and other Knowledge Bases. Impact of Social Media on TV Content Consumption - New Market Strategies, Scenarios and Trends. Proceedings of eChallenges 2015 e-2015.

Jaynes, E. T. 2003. Probability theory: The logic of science. Cambridge university press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 2014. Bayesian data analysis. Boca Raton, FL, USA: Chapman & Hall/CRCBessière, P., Mazer, E., Ahuactzin, J. M., & Mekhnacha, K. 2013. Bayesian programming. CRC Press.