Chapter 10

Cross-lingual linking of multi-word entities and language-dependent learning of multi-word entity patterns

Guillaume Jacquet
European Commission, Joint Research Centre, Ispra, Italy

Maud Ehrmann
Swiss Federal Institute of Technology in Lausanne (EPFL) – Digital Humanities Laboratory

Jakub Piskorski
European Commission, Joint Research Centre, Ispra, Italy

Hristo Tanev
European Commission, Joint Research Centre, Ispra, Italy

Ralf Steinberger
European Commission, Joint Research Centre, Ispra, Italy

We address large-scale multilingual multi-word entity (MWEntity) recognition and variant matching. Firstly, we recognise MWEntities in 22 different languages, identify monolingual variant spellings and link equivalent groups of variants across all languages. We then use the previously recognised MWEntities to learn new recognition rules based on distributional patterns. Not requiring any linguistic tools, the method is suitable for our highly multilingual environment. When adding the new rules to the original rule-based NER system, F1 performance for Spanish increases from 42.4% to 50% (18% increase) and for English from 43.4% to 44.5% (2.5% increase). Besides aiming at turning free text into semi-structured data for search and for machine-processing purposes, we use the system to link related news over time and across languages, as well as to detect trends.

Guillaume Jacquet, Maud Ehrmann, Jakub Piskorski, Hristo Tanev & Ralf Steinberger. 2019. Cross-lingual linking of multi-word entities and language-dependent learning of multi-word entity patterns. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 269–297. Berlin: Language Science Press. DOI:10.5281/zenodo.2579049

1 Introduction

Named Entities (NEs) such as persons, organisations, locations and events are major bearers of information in text as they provide answers to the text representation questions *Who did What to Whom, Where and When*. For this reason, work on NER and Classification is abundant (Nadeau & Turney 2005) and NEs have been linked to knowledge bases (Rao et al. 2013; McNamee & Dang 2009). Major challenges are homographic entity names belonging to different classes or within the same class and the existence of variant spellings within the same or across different languages, as well as morphological inflection (Steinberger et al. 2013). An additional challenge for names of organisations and events is that they may be referred to as multi-word expressions or acronyms, e.g., *Economic Community of West African States* (abbreviated as ECOWAS), and that name parts are likely to be translated, e.g., the equivalent Portuguese *Comunidade Económica dos Estados da África Ocidental* (abbreviated as CEDEAO). Users searching for such an entity will want to retrieve all mentions, independently of their spelling or abbreviation or language.

Our interest in entity variants originally stems from our multiannual work on the *Europe Media Monitor* (EMM), which is a freely accessible meta-news web platform[1] that has been online since 2002 (Steinberger et al. 2009; 2015). EMM currently gathers an average of 300,000 news articles per day in about 70 languages from about 8,000 news websites (HTML pages and RSS feeds). News items are classified into thousands of categories and related news (e.g., from different news sources) are grouped into clusters. EMM-NewsBrief and the medical information system EMM-MediSys group the newest articles every ten minutes and show intra-day trends, while EMM-NewsExplorer groups related articles published on the same calendar day and follows trends over longer periods of time. For each news article and for each news cluster, the system displays extracted meta-information, which includes the news category, entity names found (persons, organisations and geo-locations), quotations by and about entities, as well as various types of statistics, trends and analysis results. Entity mentions are disambiguated according to entity types (e.g., *Paris Hilton* is a person) and geographical reference (e.g., there are about fifteen places world-wide called *Paris*). Spelling variants of the same person or organisation name are mostly recognised as belonging to the same real-world entity. For instance, the spellings *Jean-Claude Juncker*, *Jean Cloud Junker*, *Jean-Claude Juencker*, *Жан-Клод Юнкер*, *Ζαν Κλοντ Γιούνκερ*, جان كلود جونكر, *Ζαν Κλοντ Γιούνκερ*, 让-克洛 德・容克 and many

[1] See http://emm.newsbrief.eu/overview.html and http://emm.newsexplorer.eu/

others are all identified as referring to the 12th President of the European Commission. Such multilingual entity variants – and also disambiguated place names – are a major ingredient for the successful identification of related news across languages in EMM-NewsExplorer. The system was entirely developed by the European Commission's Joint Research Centre (JRC) with the purpose of providing media monitoring functionality for the European institutions, for national authorities of the European Union (EU) Member States, for international organisations such as the United Nations or the African Union, as well as for EU partner country organisations. However, the results are also freely accessible to the wider public through web pages and as customisable mobile applications.

Person name recognition is rather well-implemented in EMM, but the coverage of multi-word organisation and event names has traditionally been rather poor because they behave like free text, i.e. they may include lower-case words, prepositions, determiners, etc. Recognising such complex MWEntity types would benefit from using syntax parsers, part-of-speech taggers, morphological analysers and generic dictionaries, but EMM cannot use these because of its need to process very large volumes of text data in near-real time and because such resources are not easily available nor quick to develop (Steinberger et al. 2013). In response to this shortcoming, the EMM team has engaged in less knowledge-intensive ways of recognising multi-word entities such as those presented in this chapter. Our general idea is to collect large numbers of known entities using patterns to recognise acronyms and their long-forms (presented in Section 3) and then to use these to learn light-weight recognition patterns for such complex MWEntities (Section 4). In order to validate this last step independently of the quality of the initially automatically created resource, we did our first experiments using MWEntity lists derived from the BabelNet resource to learn recognition patterns in a couple of languages.

In the following sections, we will first summarise the state-of-the-art for the recognition of acronyms and other multi-word entities, as well as for the recognition of monolingual and cross-lingual entity variants (Section 2). Section 3 focuses on methods and results to recognise acronyms and their expansions (e.g., *EC – European Commission*) and to identify the variant spellings and translations. In Section 4, we present different pattern learning methods that will help with the recognition of multi-word entities that are not found next to their acronyms and we will compare their relative performance. We will conclude our chapter with a summary and with pointers to future work.

2 Related work

As mentioned in the introduction, multi-word entity recognition is strongly related to acronym recognition. This statement will be further developed in the following sections.

Work in the domain of abbreviation processing is abundant, but it mostly focuses on the biomedical domain and on the English language. Since the pioneer work of Taghva & Gilbreth (1999), research has developed into three main directions, namely acronym extraction and mapping to their expansions; acronym variant clustering; and, more recently, acronym disambiguation. While the extraction of acronym/expansion pairs corresponds to the primary stage of lexical unit acquisition, variant clustering resembles sense inventory organisation, which can eventually serve as reference for disambiguation. We report here on the first two aspects.

With regard to acronym extraction, existing work almost exclusively focuses on English biomedical literature (Schwartz & Hearst 2003; Okazaki & Ananiadou 2006; James et al. 2001; Wren & Garner 2002; Adar 2004; Chang et al. 2002; Nadeau & Turney 2005). Results are good and the extraction-recognition step can be considered a mature technology for this combination of domain and language. However, there is very little work on other languages: Kokkinakis & Dannélls (2006) investigate the specificity of Swedish, Siklósi et al. (2014) carry out Hungarian abbreviation processing, both on medical texts. Kompara (2010) and Hahn et al. (2005) seem to be the only ones to work with acronyms *across* languages, with preliminary work on Slovene, English and Italian for the former, and acronym alignment across English, German, Portuguese and Spanish based on an interlingua for the latter.

As mentioned previously, the variety and the number of acronyms is very large so that it is useful to organise the acronym dataset on a semantic basis by grouping related variants under the same acronym identifier. The aim is thus – for each set of expansions having the same acronym – to identify those which are conceptually related. Previous related work focused mainly, anew, on biomedical literature in English. Adar (2004) experimented with k-means clustering based on an n-gram similarity measure and on a MeSH term similarity measure. Results showed that the n-gram based clustering performs actually better than that based on the MeSH resource. Okazaki et al. (2010) designed a more complex clustering approach, using a similarity metric based on a mixture of several features. Once the best feature setting has been acquired (by supervised machine learning), hierarchical clustering is used to induce the final variant grouping. The features used to build the similarity metric are themselves similarity measures, such as

character and word n-gram similarity. The outcome of these experiments on English abbreviations showed that character and word n-gram features contribute the most to the final result. Work on monolingual clustering of acronym variants outside the biomedical domain and for altogether 22 different languages was carried out in Ehrmann et al. (2013). Ehrmann's approach is based on hierarchical group-average clustering, where cluster homogeneity is set using an empirically determined threshold. The clustering depends on a pair-wise string similarity between expansions, using a normalised Levenshtein edit distance.

To the best of our knowledge, no work has been carried out for acronym clustering across languages. What comes closest to this or, more exactly, to its result, are multilingual lexical resources such as BabelNet (Navigli & Ponzetto 2012) or YAGO (Hoffart et al. 2013). Automatically built based on the mapping between WordNet and Wikipedia (and other resources), these resources provide (among others) multilingual variants of expansions for specific acronyms. They are inherited from cross-lingual and cross-script links provided in Wikipedia. In contrast, the work presented here starts from raw data extracted from real-life texts.

As regards learning resources for the recognition and classification of named entities and domain-specific multi-word expressions, a vast bulk of research has been reported on using weakly-supervised approaches. These are based, in particular, on the bootstrapping paradigm in which, starting from an initial set of annotated examples (or seeds), the learning process proceeds without further supervision, until a convergence criterion is reached. Some examples of the work in this field is presented in Riloff (1996); Collins & Singer (1999), and Yangarber et al. (2002).

With the emergence of large-scale knowledge bases and the availability of web-scale corpora, numerous efforts on exploiting such resources for developping named entity recognition and classification tools have been reported. For instance, Nothman et al. (2013) reports on a multilingual NER approach based on using Wikipedia links for automatically annotating a huge corpus for training purposes, whereas Downey et al. (2007) presents a novel method for detecting complex (multi-word) named entities using solely capitalisation information and n-gram statistics over a Web corpus. This approach outperformed standard supervised and semi-supervised approaches for named-entity recognition in cases of complex names of types not known in advance.

Our contribution complements prior work and focuses on exploiting the vast number of named entities contained in BabelNet (Navigli & Ponzetto 2012) for learning structurally simple and linguistically unsophisticated patterns for the recognition of multi-word named entities in various languages.

3 Creation of the multilingual MWEntity resource

In this section, we describe completed work (Jacquet et al. 2016) on recognising MWEntities and their corresponding acronyms in large volumes of text in 22 different languages, on identifying monolingual variants for the same entity and on linking the equivalent groups of variants across all languages. Figure 1 illustrates that task with an example of cross-lingual linking, which shows that we can neither assume that entities across languages have the same acronym, nor can we assume that the same acronym (within the same or across languages) refers to only one entity. The result of this work is a collection of currently 64,000 MWEntities plus their 600,000 multilingual lexical variants.

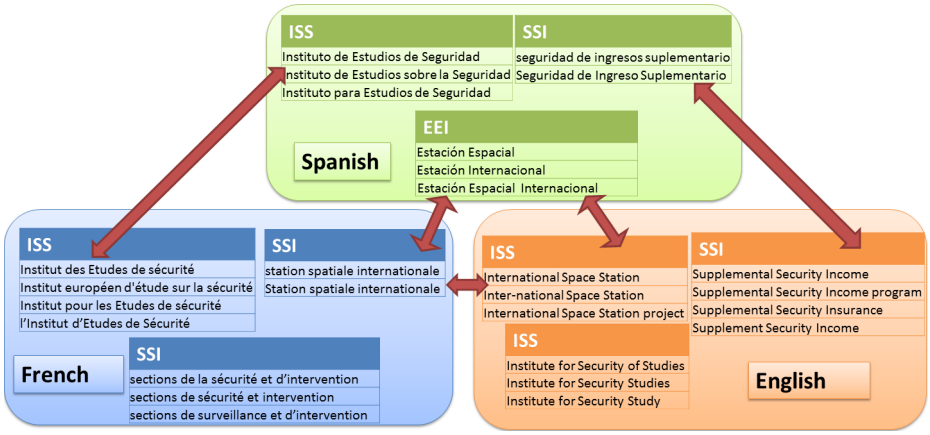

Figure 1: Example of multilingual MWEntity linking

3.1 Starting point

The starting point of our work is a large set of multi-word entities and their corresponding acronyms in 22 Roman-script languages (Ehrmann et al. 2013). These acronym/expansion pairs were extracted from the news stream analysed by the EMM processing chain by applying patterns similar to those proposed by Schwartz & Hearst (2003). In a nutshell, the algorithm collects acronym/expansion pairs (such as *expansion (acronym)* and *acronym (expansion)*) by identifying short strings within parenthesis, along with candidate expansions in a side-window of a limited length. A filtering step is then applied, with the following main constraints: the first letter of the acronym must be upper-cased, and the length of the expansion must be smaller than (a) twice as many words as there

are characters in the acronym, or (b) the number of characters in the acronym plus five words, whichever is the smaller (i.e. min(|A| + 5, |A| ∗ 2) words, with |A| being the number of characters of the acronym). We refer the reader to Schwartz & Hearst (2003) for more details. This process resulted in the extraction of 1.7 million expansions for 0.4 million different acronyms.

Applied on news articles, this method identified acronym/expansion pairs referring mostly to organisation names (e.g., *CP – Communist Party*), but also events (*WW2 – World War II*), names of drugs or of vaccines (*MMR – measles, mumps, rubella*), organisation types (*NGO – non-governmental organisation*), job titles (*MEP – Member of Parliament*), physical measurement units (*kmh – kilometres per hour*), and more. As one of the next steps, we will work on categorising the acronym/expansion pairs into various semantic categories.

To automatically determine which of the expansions are lexical variants of the same conceptual entity, a clustering step was carried out, on the basis of expansions having the same language and the same acronym. This monolingual clustering, based on a pair-wise string similarity, allowed to distinguish between sets of conceptually related expansions, such as those referring to the *International Space Station* and those referring to the *Institute for Security of Studies*, both clusters having the acronym *ISS* (cf. English part of Figure 1). Evaluated over the 10 most covered languages, this monolingual clustering has a micro-average precision of 95.2% (Jacquet et al. 2014).

Out of this monolingual clustering step, we selected only clusters having at least four expansions, resulting in 81,000 monolingual clusters with an average of 7.5 expansions per cluster, the biggest one having 232 expansions.

Based on this data, the objective is to go a step further by identifying cross-lingual multi-word entity lexical variants. More specifically, the goal is to link multilingual expansions referring to the same entity across languages and regardless of their acronyms. To this end, we leverage the previously computed monolingual clusters and attempt to link them across languages. Considering the previous example with the entity *International Space Station* (cf. Figure 1), this results in aggregating the monolingual clusters *SSI – Station spatiale internationale* (French), *ISS – International Space Station* (English) and *EEI – Estación Espacial* (Spanish). Additionally to linking expansions across languages and independently from their acronym, cross-lingual cluster aggregation can also revise monolingual clusters by aggregating those conceptually related but isolated because of their acronyms (both pairs *IMF – International Monetary Fund* and *FMI – Fondo Monetario Internazionale* occur in Italian texts).

3.2 Approach

Cluster aggregation can be cast as the problem of identifying connected components of a graph, where monolingual clusters represent vertices and where edges need to be computed. This section describes different cross-lingual aggregation strategies tested in our experiments (cf. Section 3.3) to link sets of monolingual clusters across languages.

3.2.1 Cluster aggregation based on common expansions

The most straightforward solution to link related acronyms in different languages (hereafter *ExpAgg*) is to merge those clusters that have more than n expansion forms in common, independently of whether their acronyms are identical or not (in our experiments, n was set to 1). This aggregation has been applied both to improve monolingual clusters (cf. IMF vs FMI case mentioned at the end of Section 3.1) and to aggregate clusters across languages.

3.2.2 Cluster aggregation based on tokens

3.2.2.1 Cluster representation

For the two following aggregation strategies, monolingual clusters are no longer represented by vectors of expansions, but by a vector of all individual tokens appearing in the expansions.

C is the resulting ($|\mathbb{C}| \times |\mathbb{T}|$) cluster-token matrix where $c_i : i = 1, \ldots, |\mathbb{C}|$ is a monolingual cluster, and $t_j : j = 1, \ldots, |\mathbb{T}|$ is a token. $\mathbb{T}$ contains all the tokens across languages which appear at least once in an expansion. If a token is present in different languages, such as *place* in English and *place* in French, it corresponds to different tokens in $\mathbb{T}$.

Each token has its own importance to describe a cluster. In order to compare two clusters on the basis of their most relevant tokens, we consider the tf-idf value of each token t_j where, in our context, each cluster c_i is seen as a document and the whole set of clusters $\mathbb{C}$ as a corpus:

$$C(c_i, t_j) = \mathrm{tf}(t_j, c_i) \times \mathrm{idf}(t_j, \mathbb{C}) \quad (1)$$

3.2.2.2 Cluster aggregation based on similar tokens

This aggregation (hereafter *TokAgg*) addresses cases where monolingual clusters do not have identical expansions across languages, but they have a significant amount of highly similar tokens.

Table 1: Example of clusters aggregated on the basis of similar tokens

Clusters	Expansion	Acronym	Language
cluster 1	*Social-Democratic Party* *Social Democratic Party*	*SDP*	en
cluster 2	*Partito Social-Democratico* *Partito di socialdemocratico* *Partito socialdemocratico*	*PSD*	it

We compute the matrix ($|\mathbb{T}| \times |\mathbb{T}|$), hereafter *InvEdit*, which corresponds to the inverse of the normalised Levenshtein edit distance where $t_i : i = 1, \ldots, |\mathbb{T}|$ and $t_j : j = 1, \ldots, |\mathbb{T}|$ are tokens from all the addressed languages:

$$\text{InvEdit}(t_i, t_j) = 1 - \frac{\text{Lev}(t_i, t_j)}{\max(|t_i|, |t_j|)} \tag{2}$$

$\text{Lev}(t_i, t_j)$ is the Levenshtein edit-distance between t_i and t_j, and $|t_i|$ and $|t_j|$ are respectively the length of the tokens t_i and t_j. We filter InvEdit using a threshold δ as follows:

$$\text{InvEdit}(t_i, t_j, \delta) = \begin{cases} \text{InvEdit}(t_i, t_j) & : \text{InvEdit}(t_i, t_j) \geq \delta \\ 0 & : \text{InvEdit}(t_i, t_j) < \delta \end{cases} \tag{3}$$

In this case, if $\delta = 1$, InvEdit only contains values for exact matching tokens. This matrix is then used to enrich the monolingual cluster representation. Given two languages l_1 and l_2, the corresponding monolingual clusters C_{l_1} and C_{l_2} do not have common tokens since in $\mathbb{T}$ tokens are language-dependent. The InvEdit matrix is used to identify common or similar tokens. We convert the obtained matrix $C\_\text{Tok}_{l_1}$ to a binary matrix:

$$C\_\text{Tok}_{l_1}(c_i, t_j) = \begin{cases} 1 : C_{l_1}(c_i, t_j) \times \text{InvEdit}(c_i, t_j, \delta) > 0 \\ 0 : \text{otherwise} \end{cases} \tag{4}$$

This aggregation is particularly useful when comparing clusters from similar languages. Table 1 illustrates such cases, with the English-Italian tokens *Party/ Partito* and *Democratic/Democratico*. This representation can also benefit from

the fact that it is possible to find multi-word entities of a given language in texts in another language (especially with names of international organisations such as *European Space Agency* which can be found in German text).

3.2.2.3 Cluster aggregation based on translated tokens

Table 2: Example of clusters aggregated on the basis of translated tokens

Clusters	Expansion	Acronym	Language
cluster 1	*Russian Academy of Sciences* *Russian of Academy of Sciences*	*RAS*	en
cluster 2	*russischen Akademie der Wissenschaften* *Russischen Akademie für Wissenschaften* *Russische Akademie der Wissenschaften*	*RAW*	de

However, many entities have different written forms across languages so that a string-based comparison of tokens is not successful. We therefore complement the cluster aggregation by using token translation probabilities (hereafter *TransTokAgg*).

They are produced using statistical translation models trained on parallel corpora built from Wikipedia, by making use of redirection tables (i.e. several written forms redirecting to a specific page/entity) and of interlingual links between pages (implementation details of translation models are provided in Section 3.3.3). In order to separate training and test data, any variant name from these Wikipedia tables matching with one of the 1.7 million expansions or 0.4 million acronyms is removed from the parallel corpora (see Section 3.3).

Let TransMod be the resulting ($|\mathbb{T}| \times |\mathbb{T}|$) translation model matrix where $t_i : i = 1, \ldots, |\mathbb{T}|$ and $t_j : j = 1, \ldots, |\mathbb{T}|$ are tokens. As for InvEdit matrix, we filter TransMod using a threshold β:

$$\text{TransMod}(t_i, t_j, \beta) = \begin{cases} \text{TransMod}(t_i, t_j) & : \text{TransMod}(t_i, t_j) \geq \beta \\ 0 & : \text{TransMod}(t_i, t_j) < \beta \end{cases} \tag{5}$$

This matrix is then used to enrich the monolingual cluster representation. Given a language l and its corresponding monolingual clusters C_l,C_TransTok_l corresponds to the binary extended matrix based on a given translation model:

$$C\_\text{TransTok}_l(c_i, t_j) = \begin{cases} 1 : C_l(c_i, t_j) \times \text{TransMod}(c_i, t_j, \beta) > 0 \\ 0 : \text{otherwise} \end{cases} \tag{6}$$

Table 2 illustrates a case of such cluster aggregation, thanks to a high score in the TransMod matrix between tokens *Science* in English and *Wissenschaften* in German.

3.2.3 Aggregation strategies

We formulate cluster linking as the task of identifying connected components in a graph, where monolingual clusters are vertices and where edges represent links of related clusters across languages. Clusters are linked if their similarity is above a certain threshold α. During preliminary experiments, we had also tested *pure* clustering algorithms, but it turned out that the graph approach was more efficient.

For the last two cluster aggregation methods (TokAgg and TransTokAgg), we applied two similarity measures: cosine and ComMNZ. The latter is actually a data fusion algorithm (Fox & Shaw 1994) which we assimilate, in this context, to a similarity measure. This algorithm aims at measuring the similarity between two objects having multiple comparison criteria. Specifically, the overall similarity score between two objects is better when those objects have reasonable similarity scores for all criteria than when they have a very good similarity score for one criterion, and less good or no value for the others. In our case, it would promote the similarity between two clusters c_i and c_j if they have many similar or translated tokens t_k with a reasonable similarity score, and it would decrease the similarity between two clusters c_i and c_j if they have few similar or translated tokens t_k with a very high similarity score:

$$\text{CombMNZ}(c_i, c_j) = \sum_{t_k \in c_j} \frac{C(c_i, t_k)}{\sum_{t_l \in c_i} C(c_i, t_l)} \times \sum_{t_k \in c_j} 1_{\{C(c_i, t_k) \neq 0\}} \tag{7}$$

3.3 Evaluation

3.3.1 Evaluation dataset

As described in Section 3.1, the starting point of our experiments is a set of 81,000 monolingual clusters with one acronym per cluster, an average of 7.5 expansions per cluster, many of them having few expansions, and the biggest 232 expansions.

We evaluate cross-lingual cluster aggregation against Wikipedia data excluding the part used for the translations models (see previous section). The gold standard corresponds to a set of Wikipedia redirection tables and interlingual linking tables, where we consider Wikipedia entities/pages as cross-lingual classes. Each class contains all the expressions listed in the redirection tables in all the languages linked via the interlingual linking tables. Only classes having at least two expansions were selected, resulting in a gold standard of 10,000 classes. Considering Wikipedia information as a gold standard is disputable. The interlingual linkings should be reliable but this is less the case for the redirection tables. However, a manual evaluation of the redirection table quality shows that, in over 160 randomly extracted classes in 4 different languages (fr, en, de, it), 93.4% of the forms were correct (Jacquet et al. 2014).

3.3.2 Parameters

Parameters have to be set with regards to, first, the thresholds δ and β applied to filter out some similarity values in the above-mentioned token matrices ($C\_\,\mathrm{Tok}_l$ and $C\_\,\mathrm{TransTok}_l$) and, second, the threshold α applied to the aggregation strategies, i.e. the one above which clusters are aggregated.

With respect to cluster representations based on similar tokens $C\_\,\mathrm{Tok}_l$, the threshold δ should be high in order to consider two tokens as similar only if they are close in terms of edit distance. Regarding representations based on translated tokens $C\_\,\mathrm{TransTok}_l$, the threshold β can be low since even a weak token similarity could be a relevant indicator at the cluster level. For our experiments, the values of δ and β were fixed to 0.7 and 0.3 respectively.

Cluster aggregation is allowed when the cluster similarity (either in terms of cosine or CombMNZ) is above a certain threshold α. We experimented with different values for α, ranging from 0.7 to 1 (cf. Section 3.3.5).

This aggregation step is further regulated with the addition of the following constraints: two clusters c_1 and c_2 are linked if their similarity is above α and if c_1 is in the k most similar clusters of c_2 or c_2 is in the k most similar clusters of c_1. This additional constraints allow to rule out clusters having a high similarity with a lot of other clusters. This is the case for short and frequent expansions, e.g., *Olympic Committee* which is highly similar to a cluster containing expansions such as *Olympic Organizing Committee* or to another containing *games organising committee*, but as well to clusters containing more specific expansions such as *Vancouver Olympic Committee*. In our experiments, k equals 3.

3.3.3 Translation models

Cluster representations based on translated tokens correspond to lexical conditional translation probabilities computed for three language pairs, between English and French, German and Italian. The translation models were trained on parallel corpora built from Wikipedia, by making use of redirection tables (i.e. several written forms redirecting to a specific page/entity) and of interlingual links between pages. More specifically, given an entity/page p and two redirection tables rt_1 and rt_2 in languages l_1 and l_2, each written form from rt_1 can be seen as a translation t of each written form from rt_2. For a given language pair, the corresponding parallel corpus is the concatenation of all translations t from all the entities/pages p.

These Wikipedia tables are also used for evaluation purposes (see Section 3.3.1). As a consequence, the 1.7 million expansions and 0.4 million acronyms on which the approach is applied were removed from the parallel corpora.

There were about 300,000 training examples for German–English and French–English, and about 170,000 for Italian–English. Word alignments with many-to-one links were generated using the unsupervised `fast_align` tool (Dyer et al. 2013) in both directions and combined with the `grow-diag-final-and` symmetrisation heuristic (Koehn et al. 2003). Lexical translation tables for the three language pairs in both directions where extracted with a tool from the Moses translation toolkit (Koehn et al. 2007). Tables contain maximum likelihood probability estimated for the conditional word translation probabilities $p(\text{En}|\{\text{Fr}, \text{De}, \text{It}\})$ and $p(\{\text{Fr}, \text{De}, \text{It}\}|\text{En})$. Our TransMod matrix is constructed based on the concatenation of these tables.

3.3.4 Evaluation measures

Clusters are evaluated against the gold standard using micro-average precision and recall, adopting the mapping between identified clusters and gold standard clusters which maximised the F_1 measure. Micro-average precision (MAV-P) and recall (MAV-R) are defined as follows:

$$\text{MAV-P}(C) = \frac{\sum_{c \in C} \text{EXP}(c)_{\text{true}}}{\sum_{c \in C} \text{EXP}(c)_{\text{true}} + \sum_{c \in C} \text{EXP}(c)_{\text{false}}} \tag{8}$$

$$\text{MAV-R}(C) = \frac{\sum_{c \in C} \text{EXP}(c)_{\text{true}}}{\sum_{c \in C} \text{EXP}(c)_{\text{true}} + \sum_{c \in C} \text{EXP}(c)_{\text{miss}}} \tag{9}$$

where C is the set of produced clusters, $\text{EXP}(c)_{\text{true}}$ is the set of expansions in a cluster c which also appear in the corresponding class of the gold standard, and $\text{EXP}(c)_{\text{false}}$ is the set of expansions in a cluster c which do not appear in the gold standard.[2]

3.3.5 Results and discussion

Table 3: Cluster aggregation strategies for 3 language pairs

	MAV-P	MAV-R	F1
Baseline	97.7%	51.5%	67.4%
Monolingual ExpAgg	96.8%	54.8%	69.4%
Multilingual ExpAgg	96.9%	65.7%	78.2%
Cosine measure			
TokAgg	97.7%	52.5%	68.3%
TransTokAgg	97.6%	51.8%	67.7%
All aggregations	**95.5%**	**71.4%**	**81.6%**
ComMNZ measure			
TokAgg	97.7%	52.5%	68.3%
TransTokAgg	97.7%	51.6%	67.6%
All aggregations	**95.8%**	**71.2%**	**81.6%**

Table 3 reports the results obtained for the three language pairs for which we have a translation model, and Table 4 reports on a global evaluation for 22 languages. In both cases, values were computed with the aggregation similarity threshold α set to 0.9.

We defined the baseline as the concatenation of all monolingual clusters from all languages under consideration. It has a high precision (97.7% and 98.2% in Table 3 and 4 resp.) and a poor recall (51.5% and 40.5%) since none of the clusters is cross-lingual. The challenge is thus to improve the recall without affecting the precision too much.

In Tables 3 and 4, *monolingual ExpAgg* corresponds to the expansion aggregation strategy applied at the monolingual level, and *multilingual ExpAgg* at the multilingual level. The TokAgg and TransTokAgg lines correspond to results

[2]We tried two other metrics: macro-average and B-cubed measure (Bagga & Baldwin 1998) but since results are comparable we do not report them.

Table 4: Cluster aggregation strategies on 22 languages

	MAV-P	MAV-R	F1
Baseline	98.2%	40.5%	57.4%
Monolingual ExpAgg	97.0%	44.9%	60.5%
Multilingual ExpAgg	97.4%	54.6%	70.0%
Cosine measure			
TokAgg	98.2%	45.3%	62.0%
TransTokAgg	97.7%	41.1%	57.9%
All aggregations	**93.1%**	**65.9%**	**77.2%**
ComMNZ measure			
TokAgg	98.2%	45.3%	62.0%
TransTokAgg	98.2%	40.8%	57.6%
All aggregations	**95.8%**	**65.5%**	**77.8%**

with the corresponding token aggregation strategies using cosine similarity and CombMNZ fusion, and *All aggregations* to the ones obtained when using the four aggregation strategies in a joint way.

It can be observed that each aggregation strategy contributes to improving the quality of cross-lingual cluster aggregation, with multilingual ExpAgg providing the best improvement (+10.8 points for the 3 language pairs and +12.6 points for the 22 languages). The contribution of the TransTokAgg aggregation is slightly disappointing; it improves the baseline in both language configurations, but not significantly. Nevertheless, when all the aggregations are applied (bold lines), results are better than the addition of each single aggregation. It could mean that the TransTokAgg aggregation provides links between clusters which are not useful in isolation, but adds relevant bridges between sets of clusters when combined with other aggregations. Besides, one should notice that between the three language pairs and the 22 languages, improvements per aggregation strategy are comparable. Similarly, results obtained based on cosine similarity and CombMNZ fusion are comparable. This strengthens the reliability of the obtained results.

Figure 2 shows the impact of the threshold α. When too low (0.7), the F1 measure can be below the baseline because too many links are established between clusters; when too high (1.0), aggregations based on similar and translated tokens are reduced to values close to zero. In between, it has a clear improvement impact.

Overall, all aggregations strongly improve the baseline by increasing the recall (+19.7 and +23.4 points resp.) with a small loss in precision (−1.9 and −2.4 points resp.). Eventually, there are 64,000 cross-lingual connected clusters across languages instead of 81,000 monolingual ones for the 22 languages.

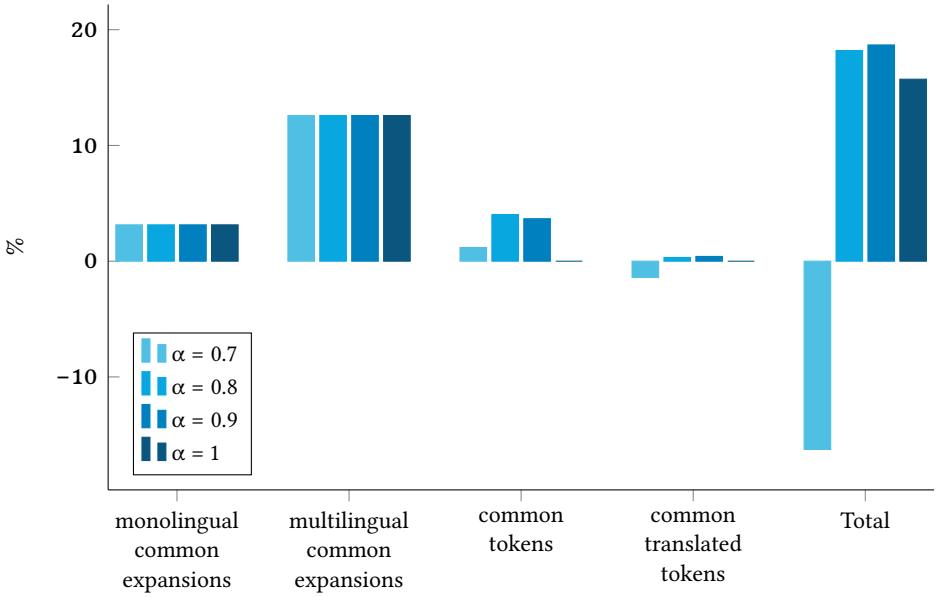

Figure 2: F1 improvement per aggregation type on 22 languages given α, using cosine similarity

4 Multi-word entity pattern learning

The previous section describes an approach which is useful to recognise frequently mentioned MWEntities and cluster them across languages, but it is limited to MWEntities mentioned at least once followed or preceded by its corresponding acronym. In this section we focus on a complementary approach to address the MWEntity recognition task. From the automatically obtained resource composed of 64,000 entities and their 600,000 multilingual lexical variants, we aim at learning MWEntity patterns in order to recognise new and not previously mentioned MWEntities. The described approach is ongoing work. Consequently, if the final goal is to learn these MWEntity patterns from the automatically extracted MWEntity resource, we must first control that our learning approach is reliable independently of the used MWEntity resource's quality. This section

describes the use of an existing and reliable resource to evaluate our pattern learning approach.

4.1 Extraction of organisation names from BabelNet

For the sake of learning multi-word entity extraction patterns we have exploited BabelNet (Navigli & Ponzetto 2012), a large multilingual encyclopedic dictionary and semantic network, created by merging various publicly available linguistic resources, e.g., WordNet and Wikipedia. In particular, BabelNet contains circa 7.7 millions of named entity-related (NE-related) synsets. We used the BabelNet API[3] to extract organisation names for English and Spanish, which were then used in the process of learning patterns in various ways. Since the NE-related BabelNet synsets are not tagged with a specific NE tag, the NE type was inferred through utilisation of the hypernym information provided in BabelNet (i.e., using WordNet hypernyms and Wikipedia categories). To be more precise, based on hypernym frequency information for the entire set of named entities a list of *positive* (circa 200) and *negative* (circa 20) hypernyms was manually created. These lists were subsequently used to extract organisation names, i.e., a given synset was extracted if: (a) there was at least one hypernym for the main sense of the synset in the list of positive hypernyms, and (b) no hypernym for the main sense of the synset was on the list of negative hypernyms. For instance, the list of positive hypernyms for extracting organisations includes terms like: *airline*, *enterprise*, *corporation*, *bank*, *local_government*, *political_organisation*, *law_enforcement_agency*, whereas the list of negative hypernyms includes terms like *person* and *human*. The main drive behind the usage of negative hypernym list was to filter out potentially ambiguous named entity candidates. In total, we have extracted 647,898 and 127,264 organisation names for English and Spanish respectively. We exploited only names that consisted of at least two tokens for the multi-word organisation name pattern learning, which resulted in maintaining only 87.0% (557,841) of the English and 86.1% (127,264) of the Spanish organisation names extracted. Noteworthy, the resource for English obtained in this manner includes a portion of organisation names in foreign languages, which is most likely due to the fact that some non-English name variants have been tagged in BabelNet as variants in English. Since the entire procedure for pulling out organisation names from BabelNet is automated such language-specific name variants have not been manually removed.

[3]http://babelnet.org/guide

4.2 Learning multi-word entity patterns based solely on BabelNet resources

The first approach to learning multi-word organisation name extraction patterns exploits as the only resource the organisation names extracted from BabelNet (see Section 4.1). Therefrom, simple linear patterns are learned that consist of two types of elements, namely, surface forms (as they appear in the organisation names) and generic token classes, which will be referred to as token class elements. Example (10) illustrates the syntax of the patterns.

(10)
```
University [] of [] the [] [UPP_W] [] in [] [UPP_W]
[ALLCAP] [] [UPP_W] [] Construction [] Group
The [] [NUM_LET] [] Company
[UPP_W] [DASH] Institute
```

`[]` denotes a whitespace (not necessarily required to be included in the pattern as illustrated by the last pattern), whereas other token classes are delimited using square brackets. There are 28 generic token classes, out of which 8 cover natural language words (e.g., `[UPP_W]` – uppercase word, `[LOW_W]` – lowercase word, `[ALLCAP]` – all capital words), letters (e.g., `[SINGCAP]` – single capital letter), numbers (e.g., `[NUM]`) and combinations thereof (e.g., `[NUM_LET]` – sequence of digits followed by a sequence of letters, etc.), whereas the remaining 20 classes are used to denote specific symbols (e.g., brackets, commas, dots, colons, etc.).

The pattern learning process consists of three main steps: (a) acquisition of candidate patterns, (b) filtering unreliable and ambiguous candidate patterns, and (c) ranking patterns. These are described in more detail below.

4.2.1 Acquisition of candidate patterns

First, each organisation name is transformed into a candidate pattern, i.e., each token which can be found in a set of predefined surface forms (consisting of keywords that trigger organisation names, e.g., *University*, and frequently occurring word forms, e.g., prepositions) remains unchanged, whereas all other tokens are mapped into a corresponding generic token class. Each candidate pattern must contain at least one surface form and at least one token-class element, otherwise it is discarded.

The set of predefined surface forms has been computed automatically and consists of word uni-grams that fulfill the following criteria: (a) it appears more than $\phi = 20$ times as part of an organisation name, (b) it does not appear on a list of known toponyms,[4] (c) it does not appear on the list of known first names and

[4]We used GeoNames resource at: http://www.geonames.org for this purpose.

surnames,[5] and (d) it is not an adjective (unless it appears very frequently). For instance, for the subset of English organisation names consisting solely of company names, the 10 top-most frequent word uni-grams that fulfill the aforementioned criteria are: *Company*, *and*, *of*, *The*, *Group*, *Corporation*, *Bank*, *de*, *Limited* and *Air*.

4.2.2 Filtering candidate patterns

In the subsequent step, a candidate pattern is discarded if:

1. its final element is the token class `[LOW_W]` (any lowercase word), or
2. it contains only surface forms which are single uppercase letters and it does not contain any token-class element representing words starting with an uppercase letter (e.g., `[UPP_W]`, `[ALLCAP]`), or
3. it starts with an initial uppercase letter, followed by an optional dot and a sequence of token classes corresponding to words starting with uppercase letters (and variations of this pattern), e.g., the following candidate pattern would be discarded: `A [] [DOT] [] [UPP_W] [] [ALLCAP]`

The filtering rules 1–2 are used in order to eliminate unreliable patterns, i.e., ones that are likely to overgenerate, whereas the filtering rule 3 aims at eliminating candidate patterns that are likely to match person names. The application of the filtering resulted in maintaining 47,496 (12,966) extraction patterns for English (Spanish), where 32.3% (41.9%) of these patterns were observed more than once. Interestingly, only 0.57% of English and 0.35% of the Spanish patterns occur more than 100 times.

4.2.3 Ranking patterns

In the final step candidate patterns are ranked with respect to their reliability based on the following general assumptions related to their structure:

- a pattern that contains either: (a) a larger fraction of surface forms vis-a-vis token-class elements, or (b) longer sequences of consecutive surface forms is deemed more reliable,
- a pattern whose final element is a lowercase surface form is deemed less reliable,

[5] We used for this purpose the *JRC Name Variant Database* and a huge list of first names extracted from Piskorski et al. (2011).

- a pattern that contains either: (a) a larger fraction of token-class elements representing single capital letters and lowercase words, or (b) longer sequences of consecutive token-class elements representing lowercase words is deemed less reliable.

The formal definition of the reliability score (Rel(p)) for a pattern p is given below, where the expressions starting with # denote the number of elements in the pattern of a specific type[6] and $\alpha = 0.2$, $\beta = 0.2$, $\gamma = 0.2$, $\delta = 0.15$, $\lambda = 0.1$ and $\kappa = 0.15$ are weighting coefficients for the various criteria used in the reliability ranking, whose values have been set based on empirical observations.

$$\begin{aligned}
\text{Rel}(p) \quad = \quad & \frac{\#\text{SurfaceForms}(p) \cdot \alpha + \#\text{ConsecutiveSurfaceForms}(p) \cdot \beta}{\#\text{NonWhitespaces}(p)} \\
- \quad & \frac{(\#\text{LowerCTokens}(p) \cdot \gamma + \#\text{ConsecutiveLowerCTokens}(p) \cdot \delta)}{\#\text{NonWhitespaces}(p)} \\
- \quad & \frac{\#\text{SingleCapitalLetterTokens}(p) \cdot \lambda}{\#\text{NonWhitespaces}(p)} \\
+ \quad & \gamma + \delta + \lambda + (1 - \text{LastElementIsLowerCToken}(p)) \cdot \kappa
\end{aligned}$$

A few examples of patterns with various reliability scores (provided in brackets) are given in (11).

(11)
```
Ministry [] of [] Foreign [] Affairs [] of [] [UPP_W]          (0.97)
Institute [] of [] [UPP_W] [] Studies                          (0.95)
[UPP_W] [] [UPP_W] [] [ALLCAP] [] at [] [UPP_W] [] University  (0.67)
St [DOT] [] [LOW_W] [] [LOW_W] [] [LOW_W] [] [LOW_W] [] school (0.24)
[UPP_W] [] [LOW_W] [] [LOW_W] [] [LOW_W] [] committee          (0.22)
```

Figure 3 depicts the distribution of patterns with respect to their reliability scores.

[6] #LowerCTokens(p) denotes the number of lowercase tokens, while #NonWhitespaces(p) denotes the number of elements in the pattern which are not whitespaces, i.e., it is a count of surface forms and token-class elements. LastElementIsLowercaseToken(p) denotes a function which returns 1 in case the last element of the pattern is a lowercase token class or 0 otherwise.

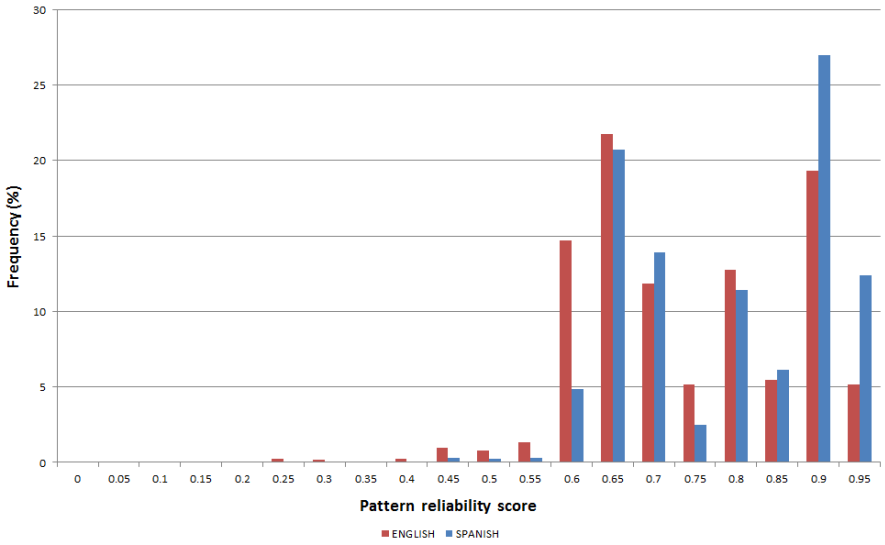

Figure 3: Distribution of patterns with respect to their reliability score for English and Spanish. Each of the bars represents the fraction of patterns, whose reliability score is within the range $(x, x + 0.5)$, where $x \in \{0, 0.5, \ldots, 0.95\}$.

4.3 Evaluation

In order to evaluate the quality of the learned patterns we used two NE annotated corpora from the CoNLL shared task for English and Spanish[7], which contains respectively 6,300 and 7,389 organisation occurrences, and tested the performance using different settings. In particular, we compared three settings: (a) using only BabelNet-derived patterns (denoted in the figures with PATTERNS), (b) using only an existing rule-based NER system as a baseline (denoted RULES), (c) combining rule-based NER system and BabelNet-derived patterns (denoted in the figures with RULES+PATTERNS). In our experiments, we used an in-house rule based NER system (Steinberger et al. 2011; Ehrmann et al. 2015) that is geared towards high precision. The choice of a specific NER system is not decisive for these experiments, but by combining an existing NER system with our BabelNet-derived patterns, we aim at testing how our automatically created patterns could be useful to improve the quality of the NER recognition.

Figures 4, 5, 6 and 7 depict the performance of applying the BabelNet-derived patterns for English and Spanish in terms of precision and recall. The precision

[7]https://www.clips.uantwerpen.be/conll2002/ner/ and https://www.clips.uantwerpen.be/conll2003/ner/

and recall values were computed for the varying minimum pattern reliability threshold in the range of $\{0.10, 0.15, \ldots, 0.95\}$, i.e., patterns below the minimum reliability threshold were discarded. The figures are showing the results obtained for exact matching (denoted with EXACT-PATTERNS or EXACT-RULES+PATTERNS) and for fuzzy matching, e.g., when there is a matching but with a left or right boundary mismatch (denoted with FUZZY-PATTERNS or FUZZY-RULES+PATTERNS). We did not visualise the results corresponding to the RULE setting in the figures because they do not depend on the reliability threshold. However, the obtained scores for this setting are embraced in Table 5.

Table 5: Results obtained with pattern reliability threshold = 0.60

	EXACT matching			FUZZY matching		
	P	**R**	**F1**	**P**	**R**	**F1**
Spanish						
PATTERNS	63.1%	10.2%	17.6%	81.4%	13.2%	22.8%
RULES	**79.8%**	24.2%	37.1%	**91.3%**	27.6%	42.4%
RULES+PATTERNS	73.5%	**31.0%**	**43.6%**	84.2%	**35.5%**	**50.0%**
English						
PATTERNS	48.5%	11.4%	18.5%	69.2%	16.3%	26.4%
RULES	**69.0%**	25.5%	37.3%	**80.4%**	29.7%	43.4%
RULES+PATTERNS	55.9%	**28.7%**	**37.9%**	65.6%	**33.6%**	**44.5%**

Table 5 provides the results obtained with a pattern-reliability threshold equal to 0.60, which corresponds to the best results obtained in both languages. For the EXACT evaluation, an improvement in terms of F1 of 6.5 points for Spanish could be observed with the setting RULES+PATTERNS versus the baseline RULES. As regards FUZZY evaluation, one could observe an improvement of F1 of 7.6 and 1.1 points for Spanish and English respectively when comparing RULES+PATTERNS versus the baseline RULES setting.

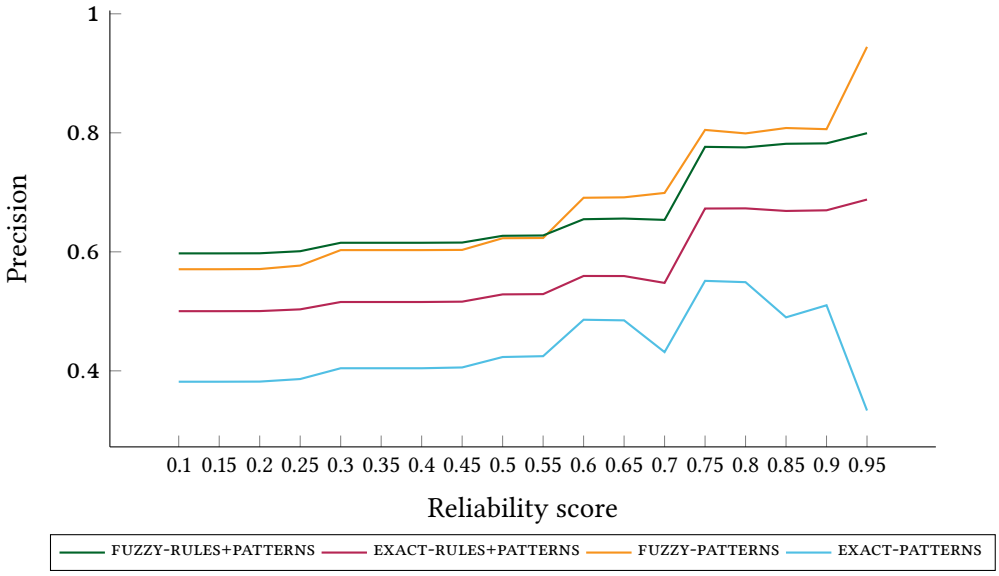

Figure 4: Experiments on English: Precision curves reflecting the performance of applying BabelNet-derived patterns and combining them with a rule-based NER system

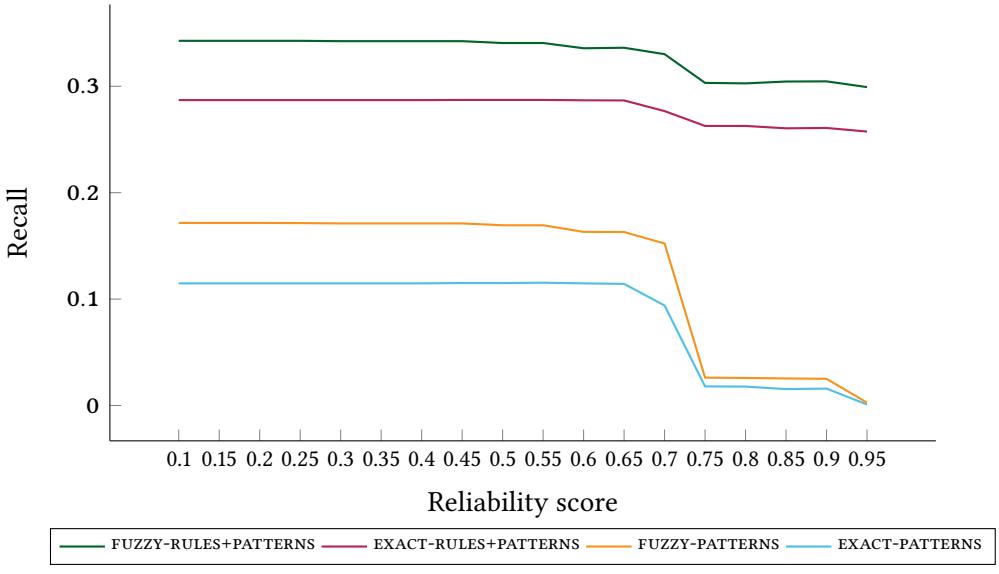

Figure 5: Experiments on English: Recall curves reflecting the performance of applying BabelNet-derived patterns and combining them with a rule-based NER system

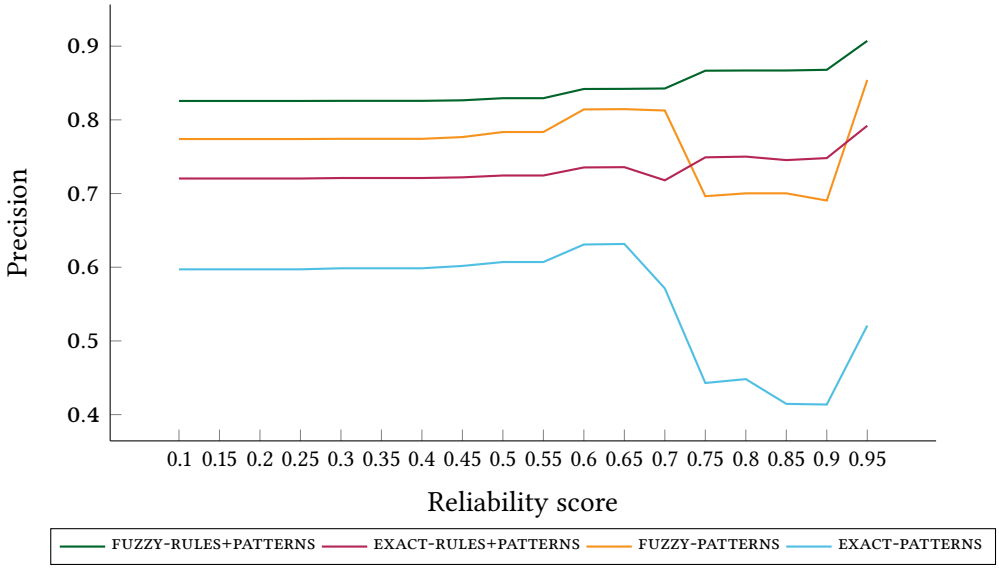

Figure 6: Experiments on Spanish: Precision curves reflecting the performance of applying BabelNet-derived patterns and combining them with a rule-based NER system

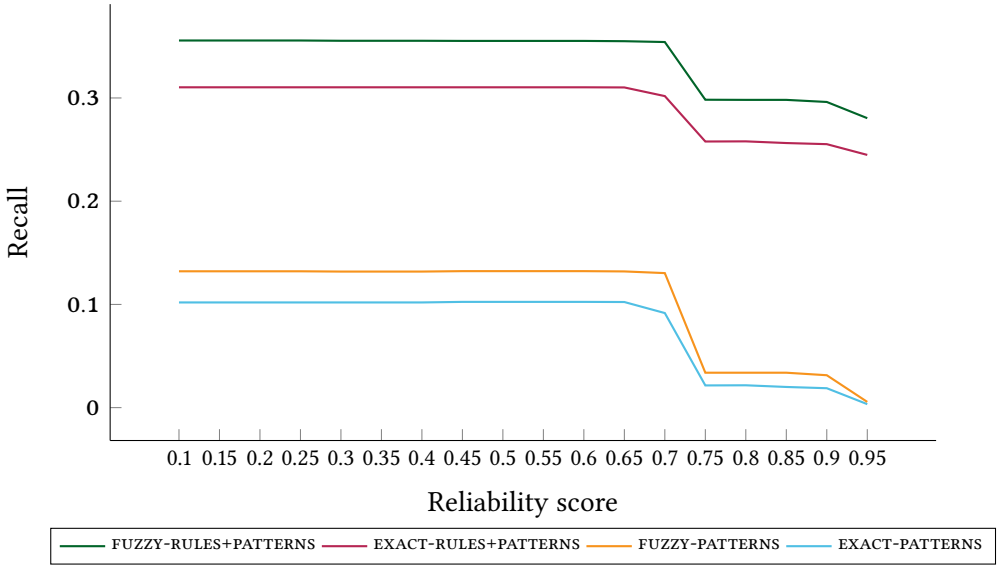

Figure 7: Experiments on Spanish: Recall curves reflecting the performance of applying BabelNet-derived patterns and combining them with a rule-based NER system

4.4 Mistakes and fuzzy matchings

Using existing NE annotated corpora for our preliminary experiments was the most obvious choice to measure the quality of our patterns. Nevertheless, it appears that some MWEntities recognised by the BabelNet-derived patterns considered as incorrectly extracted according to the annotated corpora, could still be considered as correct extractions. Figure 8 provides the complete list of "wrong" MWEntities (first column) recognised by our patterns when pattern reliability threshold equals 0.90. Even if some cases are clear mistakes, like *Results of European Super* or *Bank on Thursday*, a large fraction of the extractions could be considered as valid organisation names. The two other columns show the partial matches with the same reliability threshold, which are considered as incorrect in the *Exact matching* evaluation, and correct in the *Fuzzy matching* evaluation. Again, if some of them are clear mismatches, like *European Commission on Wednesday* or *NATO and the European Union*, most of the extractions appear to be consistent entities.

We expect to achieve higher precision of the learned patterns through embracing in the computation of the reliability score additional external evidence, i.e., exploiting contextual information obtained from pattern matchings in a web-scale corpus to judge the correctness.

"Wrong" patterns	LEFT boundary problems	RIGHT boundary problems
Major League Baseball	Brazilian Foreign Ministry	American University in Washington
Union National Bank	Chong Hing Investment Ltd	Association for Relations Across
Bank on Thursday	County Board of Education	Association of Lloyd
Results of European Super	Court of Appeals for the Fourth Circuit	Awami League of Prime
Sachs and Co.	Court of Appeals for the Ninth Circuit	Bosnian Association for Refugees
Results of European Cup	Credit Corp of USDA	British Securities and Investments
Bank of Pakistan	Democratic Party of Iran	Bureau of Alcohol
Foreign and Commonwealth Affairs	Department of Humanitarian Affairs	Department of Humanitarian
The News Agency	Gas Chemical Co Ltd	DLJ Merchant Banking Partners
Hospital School of Medicine	Illinois University at Carbondale	Economic Community of West
Police in Malta	Information Systems Co Ltd	European Commission on Wednesday
Brooks Investment Corp.	Life Insurance Co	European Union and the United States
Venture Partners of Menlo	Lodge of the University of the Witwatersrand	Government of National Unity
Department of Health	NATO and the European Union	Institute for Human Gene
Bank on Monday	Patriotic Union of Kurdistan	Irish Department of Enterprise
New England Telecommunications Corp	Rangoon Karen National Union	Kingston Technology Co.
Scottish League Cup	Rights Action League	Life Insurance Co
Police in Berovo	Staff Union of Universities	Lockhart Industries Inc.
Legislative Council on Wednesday	Taylor Made Co Ltd	Mallinckrodt Group Inc.
Yemeni Ministry of Petroleum	The Bank of Finland	Morgan Stanley and Co
	The Bank of France	National Council of Christian Churches
	The Foreign Ministry	NATO and the European Union
	The Lebanese Association for the Democracy of Elections	Security Council on Friday
	The National Basketball Association	Security Dynamics Technologies Inc
	The Republican Party	Shanghai Posts and Telecommunications
	Turkish Foreign Ministry	State National Bank
		Ukrainian Academy of Agrarian
		United in Turin
		University of Oklahoma.
		University of Pennsylvania Medical

Figure 8: Complete list of "wrong" MWEntities and left or right boundary mismatching with pattern reliability threshold = 0.90

5 Conclusion and future work

The methods described in this chapter have produced a large 22-language resource containing multi-word entities of different types and a number of automatically learned patterns to recognise newly occurring MWEntities. We intend to integrate these recognition patterns, together with the variant matching techniques, into the workflow of the *Europe Media Monitor*. An interesting feature of this collection and the patterns is that all MWEntity forms were found in real-world text and that large numbers of variants were identified, including typos, simplifications of longer names, syntactic and morphological variants and translational equivalences.

The results obtained in MWEntity recognition with the patterns automatically derived from BabelNet are promising when applied to English and Spanish, although the reported approach and evaluation figures reflect only our preliminary research in this area. Expanding the pattern learning to other languages is part of our future work. We also envisage applying the same pattern learning approach from the automatically created MWEntity resource. This would require to categorise the MWEntity sets into some broad semantic classes (e.g., organisations, events, measurements, and others) which is a task we are currently working on.

Furthermore, we are also working on expanding the patterns we learned based on a distributional approach. It consists of replacing meaningful surface forms from each pattern by a cluster of surface forms that would belong to the same semantic class. In such a way, similar words like *company*, *firm*, *corporation*, etc., will be part of the same cluster because they have a high distributional similarity. Finally, the pattern reliability scoring could be extended through inclusion of additional statistics when applying the patterns on web-scale corpora.

References

Adar, E. 2004. SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics* 20. 527–533.

Bagga, Amit & Breck Baldwin. 1998. Algorithms for scoring coreference chains. *Proceedings of the First International Conference on Language Resources and Evaluation, Workshop on Linguistic Coreference*. 563–566.

Chang, Jeffrey T., Hinrich Schütze & Russ B. Altman. 2002. Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association* 9(6). 612–620.

Collins, Michael & Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, 100–110. College Park, MD: University of Maryland.

Downey, Doug, Matthew Broadhead & Oren Etzioni. 2007. Locating complex named entities in web text. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence* (IJCAI'07), 2733–2739. Hyderabad, India: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1625275.1625715.

Dyer, Chris, Victor Chahuneau & Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 644–648.

Ehrmann, Maud, Guillaume Jacquet & Ralf Steinberger. 2015. JRC-Names: Multilingual entity name variants and titles as linked data. *Semantic Web* (Preprint). 1–13.

Ehrmann, Maud, Leo D. Rocca, Ralf Steinberger & Hristo Tanev. 2013. Acronym recognition and processing in 22 languages. In *Proceedings of the 9th conference on Recent Advances in Natural Language Processing (RANLP)*, 237–244. Hissar, Bulgaria.

Fox, Edward A. & Joseph A. Shaw. 1994. Combination of multiple searches. *NIST Special Publication*. 243–243.

Hahn, Udo, Philipp Daumke, Stefan Schulz & Kornél Markó. 2005. Cross-language mining for acronyms and their completions from the web. *Proceedings of the 8th international conference on Discovery Science (DS'05)* 9. 113–123.

Hoffart, Johannes, Fabian M Suchanek, Klaus Berberich & Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In *Proceedings of the twenty-third international joint conference on Artificial Intelligence*, 3161–3165.

Jacquet, Guillaume, Maud Ehrmann & Ralf Steinberger. 2014. Clustering of multiword named entity variants: Multilingual evaluation. In *Proceedings of the 9th Language Resources and Evaluation conference (LREC 2014)*, 2548–2553. Reykjavik, Iceland.

Jacquet, Guillaume, Maud Ehrmann, Ralf Steinberger & Jaakko Väyrynen. 2016. Cross-lingual linking of multi-word entities and their corresponding acronyms. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, 528–535. Portorož, Slovenia.

James, J. Pustejovsky, José Castano, Brent Cochran, Maciej Kotecki & Michael Morrell. 2001. Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Studies in health technology and informatics* 1. 371–375.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin & Eva Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Prague, Czech Republic.

Koehn, Philipp, Franz J. Och & Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 1, 48–54.

Kokkinakis, Dimitrios & Dana Dannélls. 2006. Recognizing acronyms and their definitions in Swedish medical texts. In *Proceedings of the 5th conference on Language Resources and Evaluation (LREC)*, 1971–1974. Genoa, Italy.

Kompara, Mojca. 2010. Automatic recognition of abbreviations and abbreviations' expansions in multilingual electronic texts. In Chris Cummins, Chi-Hé Elder, Thomas Godard, Morgan Macleod, Elaine Schmidt & George Walkden (eds.), *Proceedings of the sixth Cambridge postgraduate conference in language research (CAMLing)*, 82–91.

McNamee, Paul & Hoa T. Dang. 2009. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, vol. 17, 111–113.

Nadeau, David & Peter D. Turney. 2005. A supervised learning approach to acronym identification. In Balázs Kégl & Guy Lapalme (eds.), *Advances in artificial intelligence: Canadian AI 2005* (Lecture Notes in Computer Science 3501), 319–329. Berlin & Heidelberg: Springer.

Navigli, Roberto & Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250. DOI:10.1016/j.artint.2012.07.001

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy & James Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194. 151–175. DOI:10.1016/j.artint.2012.03.006

Okazaki, Naoaki & Sophia Ananiadou. 2006. Building an abbreviation dictionary using a term recognition approach. *Bioinformatics* 22. 3089–3095.

Okazaki, Naoaki, Sophia Ananiadou & Jun'ichi Tsujii. 2010. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics* 26. 1246–1253.

Piskorski, Jakub, Martin Atkinson & Jenya Belyaeva. 2011. Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction: A preliminary study. In *Proceedings of the conference Recent Advances in Natural Language Processing*, 210–217. Hissar, Bulgaria.

Rao, D., P. McNamee & M. Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, 93–115. Springer.

Riloff, Ellen. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of thirteenth National Conference on Artificial Intelligence (AAAI-96)*, 1044–1049. The AAAI Press/MIT Press.

Schwartz, Ariel S. & Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the PAC on Biocomputing*, 451–462.

Siklósi, Borbála, Attila Novák & Gábor Prószéky. 2014. Resolving abbreviations in clinical texts without pre-existing structured resources. In *4th workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM), LREC*, 69–75. Reykjavik, Iceland.

Steinberger, Ralf, Maud Ehrmann, Júlia Pajzs, Mohamed Ebrahim, Josef Steinberger & Marco Turchi. 2013. Multilingual media monitoring and text analysis–Challenges for highly inflected languages. In *International conference on Text, Speech and Dialogue*, 22–33.

Steinberger, Ralf, Aldo Podavini, Alexandra Balahur, Guillaume Jacquet, Hristo Tanev, Jens Linge, Martin Atkinson, Michele Chinosiand, Vanni Zavarella, Yaniv Steiner & Erik van der Goot. 2015. Observing trends in automated multilingual media analysis. In *Proceedings of the symposium on New Frontiers of Automated Content Analysis in the Social Sciences (ACA'2015)*, 1–8.

Steinberger, Ralf, Bruno Pouliquen, Mijail Kabadjov & Erik van der Goot. 2011. JRC-Names: A freely available, highly multilingual named entity resource. In *Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, 104–110. Hissar, Bulgaria.

Steinberger, Ralf, Bruno Pouliquen & Erik van der Goot. 2009. An introduction to the Europe Media Monitor family of applications. In *Proceedings of the SIGIR 2009 workshop (SIGIR-CLIR'2009)*, 1–8. Boston, USA.

Taghva, Kazen & Jeff Gilbreth. 1999. Recognizing acronyms and their definitions. *International Journal on Document Analysis and Recognition* 1(4). 191–198.

Wren, Jonathan D. & Harold R. Garner. 2002. Heuristics for identification of acronym-definition patterns within text: Towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine* 41(5). 426–434.

Yangarber, Roman, Winston Lin & Ralph Grishman. 2002. Unsupervised learning of generalized names. In *Proceedings of COLING: The 19th international conference on Computational Linguistics*. Taipei, Taiwan.