Cross-lingual Linking of Multi-word Entities and their corresponding Acronyms

Guillaume Jacquet[1], Maud Ehrmann[2], Ralf Steinberger[1], Jaakko Väyrynen[1]

[1] European Commission
Joint Research Centre
Ispra, Italy
{fname}.{lname}@jrc.ec.europa.eu

[2] Ecole Polytechnique Fédérale de Lausanne
Digital Humanities Laboratory
Lausanne, Switzerland
maud.ehrmann@epfl.ch

Abstract

This paper reports on an approach and experiments to automatically build a cross-lingual multi-word entity resource. Starting from a collection of millions of acronym/expansion pairs for 22 languages where expansion variants were grouped into monolingual clusters, we experiment with several aggregation strategies to link these clusters across languages. Aggregation strategies make use of string similarity distances and translation probabilities and they are based on vector space and graph representations. The accuracy of the approach is evaluated against Wikipedia's redirection and cross-lingual linking tables. The resulting multi-word entity resource contains 64,000 multi-word entities with unique identifiers and their 600,000 multilingual lexical variants. We intend to make this new resource publicly available.

Keywords: Multi-Word Named Entity, Named Entity Cross-lingual linking, acronyms

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 Named Entity Recognition and Classification is abundant (e.g. Nadeau and Sekine (2007)) and NEs have been linked to knowledge bases (Rao et al., 2013; McNamee and Dang, 2009). Major challenges are homographic entity names belonging to different classes or within the same class, as well as the existence of variant spellings within the same or across different languages (Steinberger et al., 2011).

The situation gets even more complex for multi-word entity names because such names are usually composed of words from the common language (e.g. *Economic Community of West African States* abbreviated as ECOWAS). These common words are normally translated when referred to in different languages (e.g. in Portuguese *Comunidade Económica dos Estados da África Ocidental*, abbreviated as CEDEAO) and authors frequently abbreviate or change the multi-word forms, either out of negligence or for space reasons (e.g. in English *Economic Community of West Africa*). While one could argue that such abbreviated or newly created names are wrong, they appear daily in real documents and an information-seeking individual or system would be interested in retrieving documents in which the *intention* was to refer to the entity of interest.

In a multilingual large-scale media monitoring environment such as EMM (Steinberger et al., 2009; Steinberger et al., 2015)[1], we observe an abundant number of spelling variants for entities, including spelling mistakes (e.g. 'United Nattions' when referring to United Nations), inflections (e.g. Birleşmiş Milletler'in where the inflection suffix *'in* is added to the Turkish equivalence of 'United Nations') and variants in other scripts (e.g. Russian Cyrillic Организация Объединённых Наций). One daily media monitoring task consists in recognising new names and in determining automatically whether they are a new name or whether they might be a spelling variant of a name encountered before. We aim at addressing this task by creating a daily updated resource containing multi-word entities, their acronyms and their variants. Ehrmann et al. (2013) developed a method handling variants at monolingual level, meaning that there were separate clusters and identifiers for each language. In this paper we address this task at the multilingual level. Additionally to the complexity of the monolingual task, we have to address expression translations, increasing acronym ambiguity, and larger numbers of expressions referring to the same entity. The *ECOWAS/CEDEAO* example mentioned previously shows how the same conceptual entity can both have different variants and different acronyms in different languages. Figure 1 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)

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

refers to only one entity.
After the discussion of related work (Section 2.), we introduce the multilingual resource that forms the starting point of our experiments (Section 3.). Next, we detail the cross-lingual cluster aggregation approach, specifying cluster representations and aggregation strategies (Section 4.). We then present our experiments, discuss the results (Section 5.) and conclude with pointers to future work (Section 6.).

2. Related work

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 and 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 and Hearst, 2003; Okazaki and Ananiadou, 2006; James et al., 2001; Wren and Garner, 2002; Adar, 2004; Chang et al., 2002; David and 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 and 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 inter-lingua 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 and 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 the many cross-lingual and cross-script links provided in Wikipedia. In contrast, the work presented here starts from raw data extracted from real-life texts.

3. Starting point

The starting point of our work is a large set of multiword 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 and 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 and 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 cate-

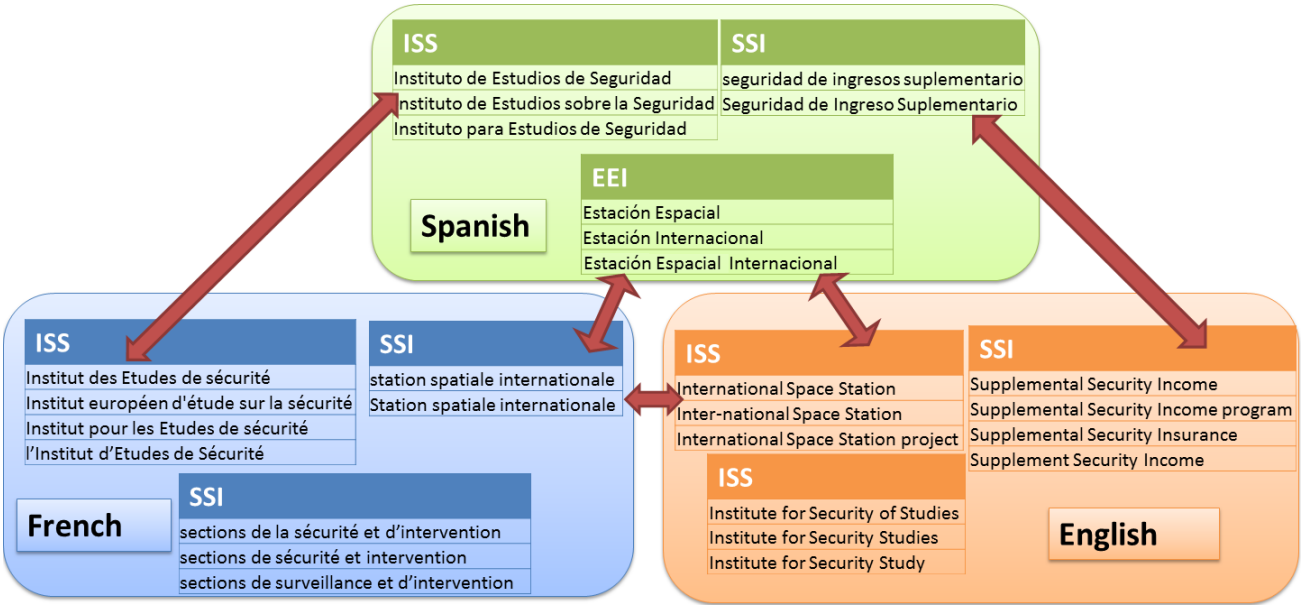

Figure 1: Example of multilingual acronym linking

gorising 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 of the current work is to go a step further by identifying cross-lingual multi-word entity lexical variants. More specifically, the objective 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).

4. 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 that we tested to link sets of monolingual clusters across languages.

4.1. 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.) and to aggregate clusters across languages.

4.2. Cluster aggregation based on tokens

4.2.1. 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 $c_i : i = 1, ..., |\mathbb{C}|$ is a monolingual cluster, and $t_j : j = 1, ..., |\mathbb{T}|$ 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 $\mathbb{T}$.

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

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

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:

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

4.2.2. 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.
We compute the matrix ($|\mathbb{T}| \times |\mathbb{T}|$), hereafter $InvEdit$, which corresponds to the inverse of the normalized Levenshtein edit distance where $t_i : i = 1, ..., |\mathbb{T}|$ and $t_j : j = 1, ..., |\mathbb{T}|$ are tokens from all the addressed languages:

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

$Lev(t_i, t_j)$ 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:
$InvEdit(t_i, t_j, \delta) =$

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

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 $C\_Tok_{l_1}$ to a binary matrix:
$C\_Tok_{l_1}(c_i, t_j) =$

$$\begin{cases} 1 : C_{l_1}(c_i, t_j) \times InvEdit(c_i, t_j, \delta) > 0 \\ 0 : otherwise \end{cases}$$

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).

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

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

4.2.3. Cluster aggregation based on translated tokens

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 5.2.1.). 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 5.).
Let $TransMod$ be the resulting ($|\mathbb{T}| \times |\mathbb{T}|$) translation model matrix where $t_i : i = 1, ..., |\mathbb{T}|$ and $t_j : j = 1, ..., |\mathbb{T}|$ are tokens. As for $InvEdit$ matrix, we filter $TransMod$ using a threshold β:
$TransMod(t_i, t_j, \beta) =$

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

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:
$C\_TransTok_l(c_i, t_j) =$

$$\begin{cases} 1 : C_l(c_i, t_j) \times TransMod(c_i, t_j, \beta) > 0 \\ 0 : otherwise \end{cases}$$

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.

4.3. 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 and 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:

$$CombMNZ(c_i, c_j) = \sum_{t_k \in c_j} \frac{C(c_i, t_k)}{\sum_{t_l \in c_i} C(c_i, t_l)} \times \sum_{t_k \in c_j} 1_{\{C(c_i,t_k) \neq 0\}} \quad (3)$$

5. Evaluation

5.1. Evaluation dataset

As described in Section 3., 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 (cf. 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).

5.2. Parameters

Parameters have to be set with regards to, first, the thresholds δ and β applied to filter out some similarity values in the token matrices (C_Tok_l and $C\_TransTok_l$) 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_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 $C\_TransTok_l$, 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 (cosine or CombMNZ) is above a certain threshold α. We experimented with different values for α, ranging from 0.7 to 1 (cf. Section 5.4.). 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.

5.2.1. 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 (cf Section 5.1.). 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` symmetrization 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 $p(En|\{Fr, De, It\})$ and $p(\{Fr, De, It\}|En)$. Our $TransMod$ matrix is constructed based on the concatenation of these tables.

	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: Cluster aggregation strategies for 3 language pairs.

	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%**

Table 4: Cluster aggregation strategies on 22 languages.

5.3. 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:

$$\text{M}-\text{AV}-prec(C) = \frac{\sum_{c\in C} EXP(c)_{true}}{\sum_{c\in C} EXP(c)_{true} + \sum_{c\in C} EXP(c)_{false}} \quad (4)$$

$$\text{M}-\text{AV}-rec(C) = \frac{\sum_{c\in C} EXP(c)_{true}}{\sum_{c\in C} EXP(c)_{true} + \sum_{c\in C} EXP(c)_{miss}} \quad (5)$$

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

5.4. Results and discussion

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 too much the precision.

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 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 *multilingualExpAgg* 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.

5.4.1. Translation model discussion

The training data for the lexical translation probabilities was quite noisy. An addition of other parallel text data might help to make more general translation tables, but it might remove some of the specificity learned from the Wikipedia data. We tried the same experiments, using the Europarl dataset,

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

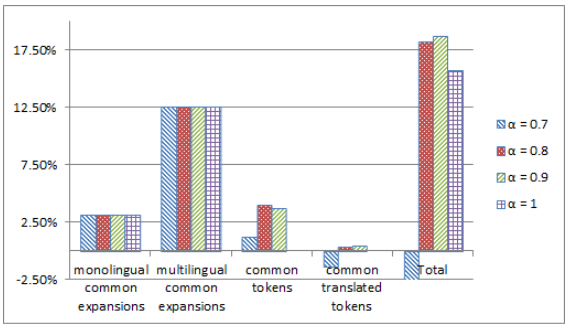

Figure 2: F1 improvement per aggregation type on 22 languages given α, using cosine similarity

but the results where comparable considering the *Cmono_TransTok* aggregation alone (F1 average = 67.5%) on the three language pairs, and the impact on the *All aggregations* was weaker (F1 average = 81.0%).

6. Conclusion

We described an approach to create a highly multilingual named entity resource consisting of acronyms and the various monolingual and multilingual spelling variants of their corresponding expansions. Thanks to different aggregation strategies, an initial set of monolingual clusters has been linked across 22 languages with a high precision (95.8%) and a reasonable recall (65.5%). The result is a resource of 64,000 unique entities with an average of 9.4 expansions (spelling variants) per cluster. Future work includes classifying the entity into types, extending the translation models to other language pairs, improving the translated token aggregation strategy, and publishing the resource as linked open data.

7. Bibliographical References

Adar, E. (2004). SaRAD: a Simple and Robust Abbreviation Dictionary. *BioInformatics*, 20:527–533.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference*.

Chang, J. T., Schütze, H., and Altman, R. B. (2002). Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Associations*, 9:262–272.

David, D. N. and Turney, P. (2005). A supervised learning approach to acronym identification. In *Proceedings of the Canadian Conference on Artificial Intelligence*.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *HLT-NAACL*, pages 644–648.

Ehrmann, M., Rocca, L. D., Steinberger, R., and Tanev, H. (2013). Acronym recognition and processing in 22 languages. In *Proceedings of the 9th Conference Recent Advances in Natural Language Processing*, pages 237–244, Hissar, Bulgaria, September.

Fox, E. A. and Shaw, J. A. (1994). Combination of multiple searches. *NIST Special Publication Sp*, pages 243–243.

Hahn, U., Daumke, P., Schulz, S., and Markú, K. (2005). Cross-language mining for acronyms and their completions from the web. *Discovery Science*, 9:113–123.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3161–3165. AAAI Press.

Jacquet, G., Ehrmann, M., and Steinberger, R. (2014). Clustering of Multi-Word Named Entity variants: Multilingual Evaluation. In *Proceedings of the 9th Language Resources and Evaluation Conference*, Reykjavik, Iceland, May.

James, J. P., Castano, J., Cochran, B., Kotecki, M., and Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Studies in health technology and informatics*, 1:371–375.

Koehn, P., Och, F. J., and Marcu, D. (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-Volume 1*, pages 48–54. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., et al. (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*, pages 177–180. Association for Computational Linguistics.

Kokkinakis, D. and Dannélls, D. (2006). Recognizing acronyms and their definitions in Swedish medical texts. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Kompara, M. (2010). Automatic recognition of abbreviations and abbreviations' expansions in multilingual electronic texts. In *Proceedings of CAMLing*, pages 82–91.

McNamee, P. and Dang, T. H. (2009). Overview of the TAC 2009 Knowledge Base Population Track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Okazaki, N. and Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–3095.

Okazaki, N., Ananiadou, S., and Tsujii, J. (2010).

Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9):1246–1253.

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

Schwartz, A. S. and Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the PAC on Biocomputing*, pages 451–462.

Siklósi, B., Novák, A., and Prószéky, G. (2014). Resolving abbreviations in clinical texts without preexisting structured resources. In *4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, LREC*.

Steinberger, R., Pouliquen, B., and van der Goot, E. (2009). An Introduction to the Europe Media Monitor Family of Applications. In *Proceedings of the SIGIR 2009 Workshop*, pages 1–8.

Steinberger, R., Pouliquen, B., Kabadjov, M., and van der Goot, E. (2011). JRC-Names: A Freely Available, Highly Multilingual Named Entity Resource. In *Proceedings of the* 8^{th} *International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, pages 104–110, Hissar, Bulgaria, September.

Steinberger, R., Podavini, A., Balahur, A., Jacquet, G., Tanev, H., Linge, J., Atkinson, M., Chinosi, M., Zavarella, V., Steiner, Y., and van der Goot, E. (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)*, pages 1–8.

Taghva, K. and Gilbreth, J. (1999). Recognizing acronyms and their definitions. *ISRI (Information Science Research Institute) UNLV*, 1:191–198.

Wren, J. D. and Garner, H. R. (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.