Clustering of Multi-Word Named Entity variants: Multilingual Evaluation

Guillaume Jacquet[1], Maud Ehrmann[2], Ralf Steinberger[1]
[1] European Commission , Joint Research Centre, Ispra, Italy
[2] Sapienza University of Rome, Italy
{guillaume.jacquet—ralf.steinberger}@jrc.ec.europa.eu, ehrmann@di.uniroma1.it

Abstract

Multi-word entities, such as organisation names, are frequently written in many different ways. We have previously automatically identified over one million acronym pairs in 22 languages, consisting of their short form (e.g. *EC*) and their corresponding long forms (e.g. *European Commission, European Union Commission*). In order to automatically group such long form variants as belonging to the same entity, we cluster them, using bottom-up hierarchical clustering and pair-wise string similarity metrics. In this paper, we address the issue of how to evaluate the named entity variant clusters automatically, with minimal human annotation effort. We present experiments that make use of Wikipedia redirection tables and we show that this method produces good results.

Keywords: Multi-Word Named Entity variants,Multilingual,acronyms

1. Introduction

Named Entities (NEs) became, over the years, one of the fundamentals of a wide variety of Natural Language Processing (NLP) applications. These lexical units, originally meant to answer the "Who did What When" of the Information Extraction paradigm, support many processes, from word alignment to event extraction, including co-reference resolution, question-answering as well as document classification, to name just a few. Within the *Europe Media Monitor* (EMM) family of applications (Steinberger et al., 2009), they are used *per se*, that is to say to retrieve information about entities of interest in news articles, as well as *metadata* to facilitate mono and cross-lingual news articles clustering.

In recent work (Ehrmann et al., 2013), we identified millions of acronym pairs in 22 different languages, consisting of a short form (SF) and its – usually several – corresponding long forms (LFs) (see Table 1 for examples). In order to automatically determine which of these multi-word entities are variant spellings of the same conceptual entity, we performed the clustering of those long forms belonging to the same short form. The clustering results need to be evaluated and, for that purpose, we propose to make use of the Wikipedia redirection tables as gold-standard variant collections.

After summarising related work (Section 2), we describe the acronym recognition method together with the acquired data (Section 3) and we present an analysis on the types of LF variants (Section 4). Next, we present our method to evaluate the NE variant clusters (Section 5), as well as the results obtained in 22 languages (Section 6). Finally, we conclude and point to future work (Section 7).

Found in English text
capital adequacy ratio
Capital Adequate Ratio
Capital Adequacy Ration
Capital Adequacy Returns
Center for Autism Research
central African Republic
Certified Automotive Recycler Program
Commission for Aviation Regulation
Confederations of Africa Rugby
Cordilleral Administrative Region
Found in French text
Caisse Autonome des Retraites
capacité africaine contre les risques
Cellule d'Action Routière
Collectif d'artistes de reggae
Collectivité d'accueil régionale
Comité d'Action pour le Renouveau
Communauté d'agglomération de Rufisque
Found in German text
Centers for Automotive Research
Central African Republic
chimären Antigenrezeptoren
Computer Assisted Reporting
Found in Italian text
Cogenerazione ad Alto Rendimento
Computer Assisted Reporting
consumo annuo di riferimento

Table 1: Multilingual examples of acronym LFs for the SF 'CAR'

2. Related work

Starting with the pioneering achievement of (Taghva and Gilbreth, 1999), much work has been completed in the domain of abbreviation processing, albeit mostly focusing on the bio-medical domain and on the English language. Research has developed into three main directions: acronym extraction and mapping to their full forms; acronym variant clustering; and, more recently, acronym disambiguation. We report here on the first two.

With regard to acronym extraction, existing approaches can be divided into four main categories ((Torii et al., 2007)): alignment-based approaches, which exploit the fact that SF and LF show letter or string ordered similarities (e.g. (Schwartz and Hearst, 2003)); collocation-based approaches, which exploit the fact that SF and LF frequently occur together ((Okazaki and Ananiadou, 2006)); pattern/rule-based approaches, which explore regularities of abbreviation conventions ((James et al., 2001), (Wren and Garner, 2002), (Adar, 2004)); and, finally, machine-

learning approaches, most of which supervised ((Chang et al., 2002), (David and Turney, 2005)). Results are good. The extraction-recognition step is a mature technology in the domain of English biomedical literature.

However, work on other languages is limited: (Kompara, 2010) describes preliminary work on Slovene, English, French and Italian. (Kokkinakis and Dannélls, 2006) investigate the specificity of Swedish. (Hahn et al., 2005) extract from English, German, Portuguese and Spanish. They are the only ones to align acronyms across languages, exploiting inter-lingua phenomena.

As illustrated in Table 1, acronyms usually exhibit a huge amount[1] and variety of lexical variants, being (typo)graphical or orthographical variants (*Capital Adequacy Ratio*, *Capital Adequacy Ration*, *capital adequacy ratio*), inflectional variants (*Clubul Alpin Roman*, *Clubului Alpin Român*) or more rarely morpho-syntactic variants (see Section 4.). Those lexical variants, once identified as such, can help organizing the acronym dataset on a semantic basis; it is therefore essential to efficiently cluster them. To this end, (Adar, 2004), working on bio-medical acronym extraction, experimented with k-means clustering based on a n-gram similarity measure first, on a MeSH term similarity measure second. MeSH based clustering allows to take into account a larger context (represented as MeSH terms) for the acronym representation, and to further validate the n-gram based one, by accepting or rejecting border line items. Results showed that the n-gram based clustering performs actually better than the MeSH based (99% accuracy rate *vs.* 76% on 555 items), for the latter has the tendency to associate semantically related acronyms, without them being lexical variants of each other. Still in the bio-medidal domain (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 is acquired (through supervised machine learning – in this case SVN classifier), hierarchical clustering is used to induce the final variant grouping. The features used to buid the similarity metric are themselves similarity measures, such as character and word n-gram similarity, Levenshtein distance, Jaro-Winkler similarity and SoftTFIDF. The outcome of those experiments showed that character and word n-gram features contribute the most to the final result; the overall performance reported was of 96% accuracy on a set of 400 abbreviations.

3. Description of the multilingual multi-word named entity data

Our data, consisting of millions of multi-word entity LFs and their corresponding SFs in 22 Roman-script languages, was extracted from the news stream analysed by the *Europe Media Monitor* (EMM; http://emm.newsbrief.eu/overview.html), which processes an average of 175,000 news articles per day in up to 74 languages. The acronym pairs, each consisting of a SF and a LF, were extracted by applying patterns similar to those proposed by (Schwartz and Hearst, 2003) for the recognition of biomedical abbreviations in English text, i.e. by identifying short strings with at least one upper-case letter in brackets (the SF) and by searching for the equivalent LF in a length-limited left-hand-side context of the bracket. At least the first letter of the SF has to be an upper-case word-initial letter. The LF must not be longer than (a) twice as many words as there are characters in the SF, or (b) the number of characters in the SF plus five words, whichever is the smaller (i.e. *min*(|A|+5, |A|* 2) words, with |A| being the number of characters of the SF). For details, see (Ehrmann et al., 2013).

Despite its simplicity, this method works astonishingly well. It successfully recognises acronym pairs such as *Namibian Broadcasting Corporation* (NBC). It usually fails to recognise foreign language acronyms such as *German Vereinte Nationen* (UNO), but it does occasionally capture foreign acronyms, as in the example *Namibische Rundfunkanstalt* (NBC). According to a manual evaluation on seven languages, the acronym recognition precision varies between 87% (French) and 98% (Hungarian), averaging around 96%. The ratio of LFs having the same SF varies between 14.67 (Latvian) and 1.98 (Basque), being 7.51 for English and 9.09 for German. The Italian variants in Table 2 give a good idea of the type of variations found.

Agenzia internazionale per l'energia atomica **AIEA** (English: IAEA)
agenzia delle Nazioni Unite per l'energia atomica
Agenzia di controllo sul nucleare delle Nazioni Unite
Agenzia internazionale energia atomica
Agenzia internazionale delEnergia atomica
Agenzia internazionale dell'Onu per l'energia atomica
Agenzia internazionale Energia atomica
Agenzia internazionale Onu per l'Energia Atomica
agenzia Internazionale per Energia Atomica
Agenzia Internazionale per il nucleare
Agenzia Internazionale per la Sicurezza Nucleare
Agenzia internazionale per l'energia atomica Onu
Agenzia nucleare delOnu
Agenzia Onu per il Nucleare
Agenzia Onu sul nucleare
Agenzia per l'Energia atomica
Agenzia per l'energia nucleare Onu
all'Agenzia internazionale dell'energia atomica
all'Organizzazione iranoana dell'energia atomica
Atomic Energy Agency
Atomica delle Nazioni Unite
dell'Agenzia dell'Onu sul nucleare

Table 2: Subset of LF variants for the Italian SF AIEA, equivalent to English IAEA - International Atomic Energy Agency. All forms were found in real-life news texts.

4. Analysis of LF variation types

In order to get an overview of the most frequent types of LF variance (e.g. different inflection forms; omission of function words; typos, etc.), we evaluated 100 LF clusters each in English and in German. We exclusively evaluated clusters of 2 LFs as the comparison of each LF with all the

[1] In previous experiments we found an average of 6.87 LFs per ambiguous – i.e. having at least 2 LFs – SF.

others in the same larger cluster would have led to results that are difficult to interpret. German was chosen as a comparison language as it uses noun compounds and nouns are weakly inflected.

For English, the majority of variations (affecting 47 of the 100 LF pairs) were due to uppercase usage or hyphenation (e.g. FCOL - flip-chip on leadframe / Flip Chip On Leadframe), simple spelling differences such as US vs. GB-English or typos (e.g. Organisation/Oganization, Amateur/Amatuer), using (or not) quotations around the LF (most such alternations could be avoided with better word tokenisation), and using abbreviations (or not) of company name parts (Corp./Corporation). The second most frequent variation phenomenon is due to the usage of morphological variants such as the genitive case (e.g. Bermuda Monetary Authority('s)), the plural form (e.g. GSC - Gulf Scrabble Championship(s)) or noun-adjective variations (e.g. CIRB - Canada/Canadian Industrial Relations Board). For 19 English pairs, the two LFs differed for more semantic reasons such as using different company designators (e.g. MCC - Mitsubishi Chemical Corporation/Company; Association/Organisation) or the omission of the company designator in one of the two LFs (e.g. Samba - Saudi American Bank (Group)). 7 pairs were affected by the addition or omission of function words (e.g. UIDC - Universal Investment (and) Development Company).

For German, which compounds two or more nouns into a single word and where nouns are weakly declined (the surface forms of several noun cases may be identical), the situation is slightly different: 41 out of 100 LF pairs differed due to the usage of morphological variants (e.g. DRF - Deutsche(n) Rettungsflugwacht); 40 pairs differed due to hyphenation (e.g. SGI - Saar-Gemeinschafts-Initiative / Gemeinschaftsinitiative), simple spelling variations or typos (Kreuzer/Kreutzer), the usage of quotes or the ampersand (using & instead of und) and of organisation designator abbreviations (e.g. Corp./Corporation); 15 LF pair differences can be described as being of a rather semantic nature such as using a name part equivalent to administration vs. organisation (HRG - Hannover Region Grundstücksverwaltung/ Grundstücksgesellschaft) or the omission of organisation designators (SFO - San Francisco International (Airport)); Finally, there were 6 cases due to the omission or usage of function words (e.g. WZB - Wissenschaftszentrum (in) Berlin).

As expected, German variants are thus more often due to the richer morphology compared to English. When checking how often more than one SF letters were drawn from a single LF word (e.g. CHAN - Championship of African Nations - both C and H were drawn from the same word championship), we found that this happened in 48 out of 200 LFs in English, but in 65 out of 200 LFs in the compounding language German (e.g. MGJ - Mädchengymnasium Jülich - girls school Jülich). Finally, while we found only one foreign language LF pair in the English set (Europäische Kommission - the German equivalent of European Commission), we found 13 out of 100 cases in German (e.g. BCV - Banque Cantonale Vaudoise, NPP - National Patriotic Party). In the German set, we also found two mixed-language LFs (BoE - Bank of England / Bank von England, ID - Intelligent Design / intelligenten Designs).

UNICEF
United Nations Childrens Fund
United Nations International Children's Emergency Fund
Unicef
U.N. Children's Fund
UN Children's Fund
United Nations' Childrens Fund
United Nation's Children's Fund
UN Children Rights
United Nations Children's Fund
Give4Free
United Nations Children's Fund
United Nation Children's Fund

Table 3: Redirect forms for the UNICEF Wikipedia page

5. Method to evaluate the named entity variant clusters

Ehrmann et al. (2013) describes a manual evaluation of the LF clusters. Despite of showing good results, this evaluation covers a limited part of the data, with only 4 languages (French, German, Italian and English) out of 22. Moreover, Recall could not be evaluated, for it would have required to look at missed LFs in all clusters. In order to enlarge the scope of this previous evaluation and to facilitate its reproducibility, we propose to use collections of variant forms extracted from Wikipedia redirection pages as gold-standard data.

5.1. Wikipedia re-direction tables as gold-standard data

Wikipedia provides dumps of re-direction tables[2] on a language basis. For each language, the redirection table contains all the forms that redirect to a specific page; as an example, table 3 shows all the forms which redirect to the UNICEF Wikipedia page. Since it is collaboratively edited and maintained by users, redirection pages potentially contain some errors. Additionally, they can include related expressions which make sense in the Wikipedia context, but which are not necessarily a lexical variant of the page title. Table 3 illustrates this case with the expression "Give4Free", the name of a programme launched by UNICEF, which however cannot be considered as a variant name for UNICEF. As a matter of fact, our gold-standard can thus contain some errors, and we need to measure their frequency.

5.2. Manual evaluation of the Wikipedia redirection data

In the context of our evaluation, each reference class corresponds to the list of forms referring to the same Wikipedia page (as in Table 3). We randomly extracted 160 classes in four different languages (French, English, German and Italian) and asked two annotators to judge the coherence of each entity of the class with respect to the most frequent one. Possible ratings were "Correct", "Too generic", "Too specific", "Related" and "Wrong". The evaluated classes

[2] http://dumps.wikimedia.org/

	Correct	Too generic	Too specific	Related	Wrong
1st annotator	91.8%	1.1%	2.7%	2.8%	1.6%
2d annotator	95%	0.4%	2.1%	1.1	1.4%

Table 4: Manual evaluation of the gold-standard.

showed an average of 3.7 forms per class. Reported in Table 4, results demonstrate that, according to the more severe annotator, most of the forms are properly clustered (91.8%), with only 1.6% considered as wrong and 6.6% as related but not corresponding exactly to the same entity. For example, *Organization for European Economic Cooperation* was rated as "too specific" compared to *Organization for Economic Development and Cooperation*. The Inter-Annotator Agreement (kappa coefficient) is 0.65, which sounds reasonable since the distribution between the different ratings is drastically asymmetric. If we reduce the rating to a binary categorisation "correct" vs "Non correct", where "non correct" corresponds to all the possible ratings except "correct", the kappa coefficient goes up to 0.74. In view of these results, *i.e.* despite the existing but quite unfrequent errors, we considered the quality of Wikipedia redirection dataset as high enough to be used as our gold-standard.

5.3. NE variant cluster evaluation against Wikipedia data

As mentioned in the introduction, our aim is not to present a new method to cluster LFs, but a new method for their *evaluation*. The clusters we evaluate were created using the method described in (Ehrmann et al., 2013) and we considered only the clusters with at least 3 LFs. Our gold-standard corresponds to the one extracted from the Wikipedia redirection tables. The evaluation was carried out using the subset of the forms at the intersection between the system dataset and the Wikipedia dataset. Table 5 shows how this filtering drastically reduce the number of evaluable LFs. From the 22 languages, 23191 LFs can be evaluated, which corresponds to 4.11% of the LFs extracted from the news.
The clusters were evaluated against the gold standard using micro-average Precision and Recall, adopting the mapping between identified clusters and gold-standard clusters that maximised the F_1 measure. For one cluster c, precision and recall are defined as follows:

$$Precision(c) = \frac{LF(c)_{true}}{LF(c)_{true} + LF(c)_{false}}, \forall c \in C \quad (1)$$

$$Recall(c) = \frac{LF(c)_{true}}{LF(c)_{true} + LF(c)_{missing}}, \forall c \in C \quad (2)$$

C corresponds to the set of produced clusters, $LF(c)_{true}$ is the set of LFs in a cluster c which also appear in the corresponding cluster of the gold-standard, and $LF(c)_{false}$ is the set of LFs in a cluster c which do not appear in the gold-standard one. Thus:

$$\text{M-AV-}prec(C) = \frac{\sum_{c \in C} LF(c)_{true}}{\sum_{c \in C} LF(c)_{true} + \sum_{c \in C} LF(c)_{false}} \quad (3)$$

$$\text{M-AV-}rec(C) = \frac{\sum_{c \in C} LF(c)_{true}}{\sum_{c \in C} LF(c)_{true} + \sum_{c \in C} LF(c)_{missing}} \quad (4)$$

With the micro-average measure, the averaging is done at the object level (here each LF is an object). We also used two other metrics: the macro-average measure, which consists of averaging at the cluster level and not at the object level, and the B-cubed measure (Bagga and Baldwin, 1998). Since the results with those metrics are comparable to those obtained with the micro-average measure, we are not reporting them in the present paper.
In order to better understand our results (described as "normalConfig" in Table 5), we compared them with different baselines: *all-in-one*, where all forms are clustered into a single cluster (leading to high Recall, but low Precision), and *singleton*, where each form is clustered as a separate single cluster (thus eading to high Precision, but low Recall). Additionally, the *randomValues* configuration evaluates the case in which the average number of LFs per cluster is the same as the "normalConfig" configuration, only that the LFs are randomly switched.

6. Experimental evaluation results for 22 languages

Table 5 reports, separately for 22 languages, how well the LF clusters match the gold-standard Wikipedia redirection data. The results look convincing since the average Precision is 95.2% and the Recall is 74.8%. Moreover, they are rather stable across all languages, as well as consistent with the manual evaluation results reported in (Ehrmann et al., 2013). In this table, all the languages with not significant enough evaluable data, i.e. having less than 100 evaluated LFs, are greyed and are not considered for the average precision and recall. Nevertheless, even if the results for these greyed languages are not significant, the values are also consistent with those for the languages with better coverage.

7. Summary and conclusion

We presented our multi-word NE dataset, consisting of millions of multi-word NEs in 22 languages, as well as our effort to automatically cluster the variant spellings into groups belonging to the same NE. We used Wikipedia redirection tables as the gold-standard ground truth with which to evaluate the resulting sets of variant spellings. The evaluation results achieved with this method are good and show that the proposed method is good enough to guarantee good quality data. Next steps include performing more significant evaluation for the less covered languages and to then tackle the next challenge, which is to automatically link variant spellings *across* languages. The final result, consisting of large numbers of entities and their many monolingual and cross-lingual spelling variants, will be distributed to the R&D community as part of the multilingual person name variant resource *JRC-Names* (Steinberger et al., 2011).

ISO lang	Language	eval config	Nb news dataset (LFs member of a cluster having at least 3 LFs)	Nb wikipedia LFs	Nb evaluated LFs	System LFs/cluster average	gold-stand. LFs/cluster average	micro-average-precision	micro-average-recall	micro-average-F1
en	**English**	**normalConfig**	**205682**	**112460**	**14080**	**1.8**	**2.90**	**96.9%**	**73.4%**	**83.5%**
		randomValues	205682	112460	14080	1.8	2.90	66.4%	34.4%	45.4%
		all-in-one	205682	112460	14080	14080.0	2.90	0.0%	100.0%	0.0%
		singletons	205682	112460	14080	1.0	2.90	100.0%	34.4%	51.2%
fr	**French**	**normalConfig**	**130803**	**36717**	**3554**	**1.8**	**2.56**	**94.7%**	**81.1%**	**87.4%**
		randomValues	130803	36717	3554	1.8	2.56	61.7%	39.1%	47.9%
		all-in-one	130803	36717	3554	3554.0	2.56	0.1%	100.0%	0.1%
		singletons	130803	36717	3554	1.0	2.56	100.0%	39.1%	56.2%
de	**German**	**normalConfig**	**46863**	**31323**	**1831**	**1.6**	**2.32**	**98.1%**	**80.1%**	**88.2%**
		randomValues	46863	31323	1831	1.6	2.32	68.0%	43.1%	52.8%
		all-in-one	46863	31323	1831	1831.0	2.32	0.1%	100.0%	0.3%
		singletons	46863	31323	1831	1.0	2.32	100.0%	43.0%	60.2%
es	**Spanish**	**normalConfig**	**87814**	**33049**	**1382**	**1.66**	**2.64**	**97.3%**	**74.4%**	**84.3%**
		randomValues	87814	33049	1382	1.66	2.64	67.8%	38.0%	48.7%
		all-in-one	87814	33049	1382	1382.00	2.64	0.2%	100.0%	0.4%
		singletons	87814	33049	1382	1.00	2.64	100.0%	37.8%	54.9%
pt	**Portuguese**	**normalConfig**	**28990**	**17572**	**642**	**1.72**	**2.53**	**96.4%**	**78.7%**	**86.6%**
it	**Italian**	**normalConfig**	**11782**	**16217**	**579**	**1.7**	**2.39**	**97.1%**	**82.0%**	**89.0%**
nl	**Dutch**	**normalConfig**	**5996**	**14694**	**313**	**1.83**	**2.25**	**100.0%**	**89.5%**	**94.4%**
ro	**Romanian**	**normalConfig**	**39358**	**4873**	**180**	**1.73**	**2.50**	**100.0%**	**81.1%**	**89.6%**
sv	**Swedish**	**normalConfig**	**1963**	**13757**	**137**	**1.99**	**2.32**	**99.2%**	**92.0%**	**95.5%**
no	**Norwegian**	**normalConfig**	**2937**	**10175**	**119**	**1.78**	**2.20**	**100.0%**	**88.2%**	**93.7%**
lt	*Lithuanian*	*normalConfig*	*7238*	*3669*	*57*	*1.68*	*2.19*	*100.0%*	*86.0%*	*92.5%*
ca	*Catalan*	*normalConfig*	*1361*	*9070*	*56*	*1.81*	*2.24*	*100.0%*	*89.3%*	*94.3%*
pl	*Polish*	*normalConfig*	*3399*	*10890*	*48*	*1.45*	*2.18*	*100.0%*	*77.1%*	*87.1%*
hu	*Hugarian*	*normalConfig*	*8090*	*4655*	*44*	*1.52*	*2.20*	*100.0%*	*79.5%*	*88.6%*
lv	*Latvian*	*normalConfig*	*8428*	*2580*	*41*	*1.95*	*2.16*	*100.0%*	*95.1%*	*97.5%*
cs	*Czech*	*normalConfig*	*3687*	*7373*	*37*	*1.54*	*2.31*	*100.0%*	*78.4%*	*87.9%*
da	*Danish*	*normalConfig*	*1601*	*6686*	*27*	*2.25*	*2.25*	*100.0%*	*100.0%*	*100.0%*
fi	*Finnish*	*normalConfig*	*1643*	*8623*	*26*	*2.00*	*2.17*	*100.0%*	*96.2%*	*98.0%*
et	*Estonian*	*normalConfig*	*2650*	*3657*	*18*	*1.64*	*2.00*	*100.0%*	*88.9%*	*94.1%*
sl	*Slovene*	*normalConfig*	*4991*	*2641*	*12*	*1.20*	*2.00*	*100.0%*	*66.7%*	*80.0%*
sw	*Swahili*	*normalConfig*	*1338*	*872*	*6*	*1.50*	*2.00*	*100.0%*	*83.3%*	*90.9%*
eu	*Basque*	*normalConfig*	*88*	*2901*	*2*	*2.00*	*2.00*	*100.0%*	*100.0%*	*100.0%*
Average		**normalConfig**	**606702**	**354454**	**23191**	**1.7**	**2.29**	**95.2%**	**74.8%**	**83.7%**

Table 5: LFs cluster evaluation for 22 languages

8. 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*.

Ehrmann, M., Rocca, L., 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.

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.

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.

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.

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.

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.

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 (SIGIR-CLIR'2009)*, pages 1–8, Boston, USA, July.

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 8th International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, pages 104–110, Hissar, Bulgaria, September.

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

Torii, M., Hu, Z., Song, M., Wu, C. H., and Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*, 8 (suppl. 9).

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.