Identification of Disease Symptoms in Multilingual Sentences: an Ontology-Driven Approach*

Angelo Ferrando[1], Silvio Beux[1], Viviana Mascardi[1], and Paolo Rosso[2]

[1]DIBRIS, Università degli Studi di Genova, Italy
angelo.ferrando@dibris.unige.it, silviobeux@gmail.com,
viviana.mascardi@unige.it
[2]PRHLT, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. In this paper we present a Multilingual Ontology-Driven framework for Text Classification (MOoD-TC). This framework is highly modular and can be customized to create applications based on Multilingual Natural Language Processing for classifying domain-dependent contents. In order to show the potential of MOoD-TC, we present a case study in the e-Health domain.

Key words: Multilingual Natural Language Processing, Ontology-Driven Text Classification, BabelNet, Symptom Disease Identification

1 Introduction

The large amount of digital data made available in the last years from a wide variety of sources raises the need for automatic methods to extract meaningful information from them. The extracted information is precious for many purposes, and especially for commercial ones. Jackson and Moulinier [12] observe that *"there is no question concerning the commercial value of being able to classify documents automatically by content. There are myriad potential applications of such a capability for corporate Intranets, government departments, and Internet publishers"*.

The problem of classifying multilingual pieces of text was addressed since the end of the last millennium [17] but it is still a significant problem because each language has its own peculiar features, making the automatic management of multilingualism an open issue.

The use of ontologies to classify multilingual texts [5] is a good alternative to standard machine learning approaches in all those situations where a training set of documents is not available or it is too small to properly train the classifier. Ontology-driven text classification does not depend on the existence of a training set, as it relies solely on the entities, their relationships, and the taxonomy of categories represented in an ontology, that becomes the driver of the

* The first author of this paper is a PhD student in Computer Science at the University of Genova, Italy. The work of the last author was in the framework of the SomEMBED MINECO TIN2015-71147-C2-1-P research project.

classification. Another advantage of ontology-driven classification is that ontology concepts are organized into hierarchies and this makes possible to identify the category (or the categories) that best classify the document's content, by traversing the hierarchical structure.

In this paper we present MOoD-TC (*M*ultilingual *O*ntology *D*riven *T*ext *C*lassifier [3, 13]), a highly modular system which has been conceived, designed and implemented to be customized by the system developer for obtaining different domain-dependent behaviors, always centered around the multilingual text classification process. The original contribution of this paper is the exploitation of the core "multilingual word identification" functionalities of MOoD-TC for a challenging scenario in the e-Health domain, where classification is a by-product of disease symptoms identification in multilingual pieces of text, driven by a standard symptoms ontology. A customization of MOoD-TC with an ad-hoc module equipped with pre- and post-processing facilities suitable for the scenarios that motivate our work, is also described.

The paper is organized as follows: Section 2 introduces three motivating scenarios where an ontology-driven multilingual text classification may prove useful, Section 3 analyzes the state of the art, Section 4 describes MOoD-TC, Section 5 provides examples and experimental results, and Section 6 concludes.

2 Motivating scenarios

Alice is enjoying her holidays in Stockholm. Suddenly, she feels a painful spasm to her stomach and in a few minutes a strong feeling of nausea appears. Spasms go on for half an hour, and she starts to feel worried. She does not think it is the case to go to the hospital, but she would at least ask for advice over the phone. However, she cannot speak Swedish and, in the stressful situation she is experiencing, she cannot recall how to express her health problems in English. She could speak in her native language Italian, but it is not so likely that the doctor can speak Italian as well.

Bob is making a walk in his town. He notices a young man bending over his knees, with a scared expression on his face. He runs to help him, and he understands that the problem is with his chest. The young man speaks French only and Bob cannot understand him: he calls the first aid emergency number and explains what he is seeing and what he supposes to be taking place. If he could understand what the young man says, he would be definitely more helpful.

Carol is a volunteer in Honduras. She is neither a physician nor a nurse. She has a very basic knowledge of first aid procedures and a first aid kit with medicines that she knows how to administer, given a clear diagnosis. A woman runs towards her asking for her assistance. The woman's small boy has a problem with his head and he has a high fever but, without understanding the other symptoms that the woman is trying to explain in Spanish, Carol cannot recognize and classify the problem. In the remote place where she is, she cannot contact the doctor. Carol should need to understand the other symptoms besides fever and headache, in order to select the correct medicine.

The three scenarios above are all characterized by the impossibility for the doctor to visit the patient on-the-fly and the need for the patient to be understood despite language barriers, in order to get advice for minor problems or to

speed up the assistance procedure for major ones. The patient's need could be suitably addressed by identifying and translating symptoms from her language to the assistant's or the doctor's one. If automatic tools for facing this issue were available, for example as an app installed on the mobile phone, the three situations could evolve in the following way:

- **Scenario 1**: through the use of an app, the person needing care communicates with the "health emergency" software application in her own language. The application performs a speech-to-text translation, **identifies the symptoms in the text based on a standard ontological representation of symptoms**, and sends the list of symptoms expressed in the doctor's language to a center where they are managed either by intelligent software agents or by human experts.
- **Scenario 2**: the "health emergency" software application is not directly used by the person needing care, but by the one who assists her. Like before, the assisted person can "tell" her problems to the application which performs a speech-to-text translation and **identifies the symptoms represented in a domain ontology which appear in the text**. The symptoms, translated into the language of the person who his giving the first assistance, may be read on the screen. That person can call the national first aid number, telling what is happening, what she sees, and the symptoms which have been understood, classified, and translated by the app.
- **Scenario 3**: also in this case, besides a speech-to-text translation, **the symptoms expressed in the language of the patient are identified w.r.t. a symptoms ontology** and translated into the target language. The way this information is used can require a further automatic processing stage, if the doctor cannot be involved in the loop and the person providing aid needs an automatic support for making a diagnosis and identifying the right therapy to administer.

In all the three situations above, a standard machine translation application and a symptoms classifier based on machine learning might not be powerful enough: the pre- and post-processing stages require to have a machine-readable explicit representation of symptoms, in some vocabulary agreed upon by all the application components and by the humans involved in the loop, in order to share them among the application components (both at the client and at the server side) and to reason about them if needed. A multilingual ontology-driven text classification approach seems the right way to face these challenging scenarios.

3 State of the art

According to [8], in 1996 more than 80% of Internet users were native English speakers. This percentage has dropped to 55% in 2000 and to 27.3% in 2010. However, about 80% of the digital resources available today on the Web (including deep Web and digital libraries) are in English [10]. This calls for the urgent need of establishing multilingual information systems and Cross-Language Information Retrieval (CLIR) facilities. How to manipulate the large volume of multilingual data has now become a major research question.

In this paper we are interested in Natural Language Processing (NLP) techniques for solving multilingual term identification and text classification problems in the e-Health domain where extracting information from clinical notes has been the focus of a growing body of research in the past years [14]. Common characteristics of narrative text used by physicians in electronic health records make the automatic extraction of meaningful information hard. NLP techniques are needed to convert data from unstructured text to a structured form readily processable by computers [15]. This structured representation can be used to extract meaning and enable Clinical Decision Support systems that assist healthcare professionals and improve health outcomes [6].

Signs and symptoms have seldom been studied for themselves in the field of biomedical information extraction. Indeed, they are often included in more general categories such as "clinical concepts" [22], "medical problems" [21] or "phenotypic information" [19]. Moreover, most of the available studies are based on clinical reports or narrative corpora. In [11, 18], indeed, the aim consists in symptom extraction from clinical records and in [20] the authors identify the risk factors for heart disease based on the automated analysis of narrative clinical records of diabetic patients.

Another recent project in e-Health NLP context is the IBM Watson for Oncology[1]. It has an advanced ability to analyze the meaning and context of structured and unstructured data in clinical notes and reports, easily assimilating key patient information written in plain English that may be critical to select a treatment pathway. These works are different from ours because they do not address multilingual aspects and, furthermore, because they have to manage the differences between the "signs", which are identified by clinicians, and the "symptoms", which can be described directly by the sick person.

In our work we do not have to manage clinical records but directly the information provided by the person who feels sick. This difference is crucial in works using an ontology-driven approach, because clinical reports contain many more technical words[2] compared to a text written (or a sentence told) by a normal person describing how she feels. This allows us to use simpler ontologies. Especially from the multilingual viewpoint, having an ontology containing simple concepts, omitting useless technicalities, allows achieving better results with less effort, considering that a technical word could be less supported by the tools we use during our text classification pipeline.

The assumption upon which MOoD-TC relies, is the availability of ontologies in the domain of interest. Even if the application developer might design and implement her own domain ontology from scratch, integrating well-founded and widely used ontologies into MOoD-TC would be the most modular, reusable and scientifically acceptable approach. Luckily, many domain ontologies exist, in particular in the biomedical domain. Panacea [7], the Ontology for General Medical Science[3], and the Gene Ontology[4] are just a few recent examples, besides the "symptoms ontology" used for our experiments and discussed in Section 5.

[1] http://www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html
[2] A clinical report is written by a doctor.
[3] https://bioportal.bioontology.org/ontologies/OGMS
[4] http://geneontology.org/

4 MOoD-TC

MOoD-TC has been developed as part of Silvio Beux' Masters Thesis [3], starting from [13]. Its aim is to classify multilingual textual documents according to classes described in a domain ontology. MOoD-TC consists of the Text Classifier (TC) and the Application Domain Module (ADM). It provides a set of core modules offering functionalities which are common to any text classification problem (text pre-processing, tagging, classification) plus a customizable structure for those modules which can be implemented by the developer in order to offer application-specific functionalities. It returns a classification of the text w.r.t. the ontology taken as input. The classification performed by TC which is implemented in Java and exploits the Language Detector Library[5], BabelNet [16], and TreeTagger[6].

The Language Detector Library detects, with a precision greater than 99%, 53 languages making use of Naive Bayesian filters. It is devoted to recognize the language L_o of the ontology o and the language L_d of the textual document d. The TreeTagger tool performs the tagging of d in order to obtain, for each word $w \in d$ different from a stop word, its lemma (the canonical form of the word) and its part of speech (POS). This information is used by BabelNet to perform the translation of w into the ontology language. Finally, the translated word w' is searched inside the ontology and contributes to the classification of d in the category modeled by the ontology concept c having the same semantics as w'. The *ClassifierObject* is the object that stores a correctly classified word (and additional information) of the document d with respect to o. TC returns a list of such objects. ADM specializes the text classifier task by implementing

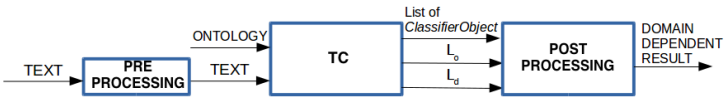

Fig. 1. Integration pipeline of TC and ADM.

functionalities for pre- and post- processing a multilingual textual document. If an ADM is used, the entire system specializes its behaviour in the domain represented by that particular ADM (e.g., from text classifier to disease recognizer). In our system TC can work alone, but an ADM is meant to work in close connection with the core system. The core modules are implemented to work for the European languages (which share some common features like, for example, the relationship between noun and adjective), but they could be extended to cope with the peculiar features of other languages; in fact, thanks to the modularity of the system, it is possible to integrate different algorithms created specifically to handle that peculiarities, without modifying the entire system. The ADM processes the TC input and output in order to obtain a new domain oriented tool. An ADM is composed by two sub-components: pre-processing and post-processing. The pre-processing component takes as input a digital object (for

[5] https://code.google.com/p/language-detection/
[6] http://code.google.com/p/tt4j/

example a spoken sentence, in the scenarios discussed in Section 2) and returns a new processed text, while the post-processing component takes as input the TC output and returns a domain dependent result. Figure 1 shows the entire pipeline of the integration process between the TC and the ADM.

5 Exploiting MOoD-TC for Symptom Identification

As illustrated in Section 2, the scenarios we aim to address require that disease symptoms appearing in a text are correctly identified w.r.t. a domain ontology. The pre-processing stage consists of moving from a spoken sentence to a text and the post-processing in translating the identified symptoms into a target language and, depending on the scenario, moving back from text to speech and/or reasoning over them. In the sequel we discuss the experiments related with our main task, namely that of symptoms identification.

The domain ontology used for describing symptoms is a standard ontology named the *symptoms ontology*[7], partially shown in Figure 2. It is an ontology of disease symptoms with symptoms encompassing perceived changes in function, sensations or appearance reported by a patient and indicative of a disease. We stress that our experiments in exploiting MOoD-TC for symptom identification did not require to build any new ontology. Rather, consistently with the good principle of reusing existing software whenever available and, in particular, reusing existing ontologies, we just passed the symptoms ontology as input to the TC, obtaining the results discussed in the next section.

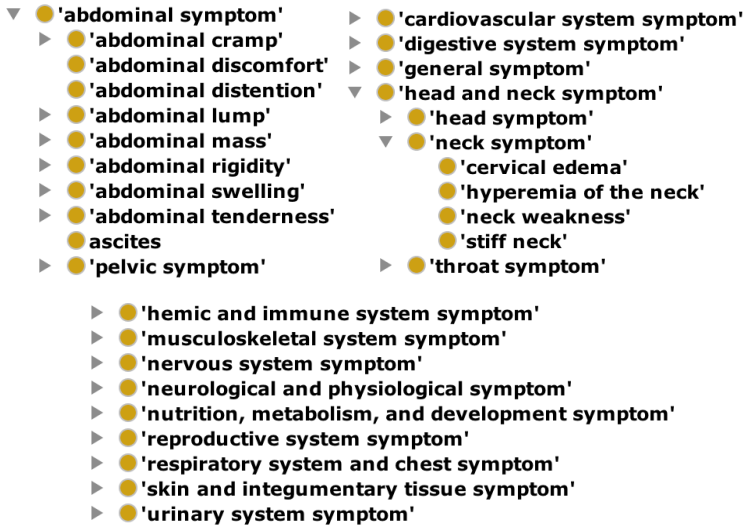

Fig. 2. Symptoms ontology (the three branches are children of "Symptom").

In the sequel we discuss our initial experiments with phrases in five different languages (English, French, German, Italian, Spanish), where symptoms are

[7] http://purl.obolibrary.org/obo/symp.owl

identified by the TC module. The classification of two sample sentences is shown below, where the TC GUI screenshot associated with each sentence shows the ontology concepts which appear in the text along with the number of their occurrences in the text.

Phrase 1 (Italian language): *"Credo di avere la febbre, continuo a sudare e ho i brividi. Non la smetto di tossire e fatico a mangiare a causa del male alla gola, come un forte bruciore. Mi sento stanchissimo e ho dolore a tutti i muscoli."*

Lemma word	Ontology word	Occurences
febbre	fever	1
brivido	tremor	1
tossire	cough	1
male gola	pain_throat	1
dolore muscolo	pain_muscle	1

Phrase 3 (Spanish language): *"Me siento fatal. Tengo temperatura, vòmito y diarrea. Hace dos dìas que no consigo comer nada. Tengo nausea y mareos."*

Lemma word	Ontology word	Occurences
temperatura	fever	1
vómito	vomiting	1
diarrea	diarrhea	1
nausea	nausea	1
mareo	dizziness	1

The experiments have been carried out on 32 sentences for each of the 5 languages, for a total of 160 sentences. Each sentence describes symptoms related to one of the following sixteen disease: tinnitus, food allergy, cervical, dehydration, hyperthyroidism, flu, appendicitis, food poisoning, labyrinthitis, narcolessia, pneumonia, diabetes type 1, hyperglycemia, hypoglycemia, bronchitis, jet lag (two sentences for each disease). To cover the widest range of cases we considered the diseases with the most varied symptoms. The description of symptoms associated with each disease has been retrieved from [9] and each sentence contains 2 up to 9 symptom words. The sentences were manually created by the authors.

Since the final purpose of this work is to provide an automatic diagnostic system with as many symptoms as possible, in order to devise the correct diagnosis, we were mainly interested in symptoms which appear in the text but which are not identified by our classifier (false negatives). We also looked for false positives, but their number is so low to be irrelevant for our experiments. Also, false positives are due to an under classification, rather than an actually wrong classification: if the text contains the "abdominal cramp" symptom, for example, and it is classified with the more general "abdominal symptom" concept, we consider this result a false positive as a more specific concept could have been returned. Figure 3 shows the average number of symptoms that should have been identified w.r.t the correctly identified symptoms in the five considered languages. Figure 4 shows the number of false negatives (y axis) for disease (x axis). Figure 3 demonstrates that the results greatly vary with the disease. For example, symptoms related to tinnitus are hardly classified, but this can be easily explained by observing the ontology we used, where problems related to ears are not modeled at all. By carefully analyzing the obtained results, we also realized that sometimes the performances of the classifier are worsened by the presence of a symptom in the text which has a different grammatical role than the symptom in the ontology (usually a noun), making their matching impossible although the word root and the meaning are the same. For example, the ontology contains the noun "irritability", but if the text contains the adjective "irritable" (in any

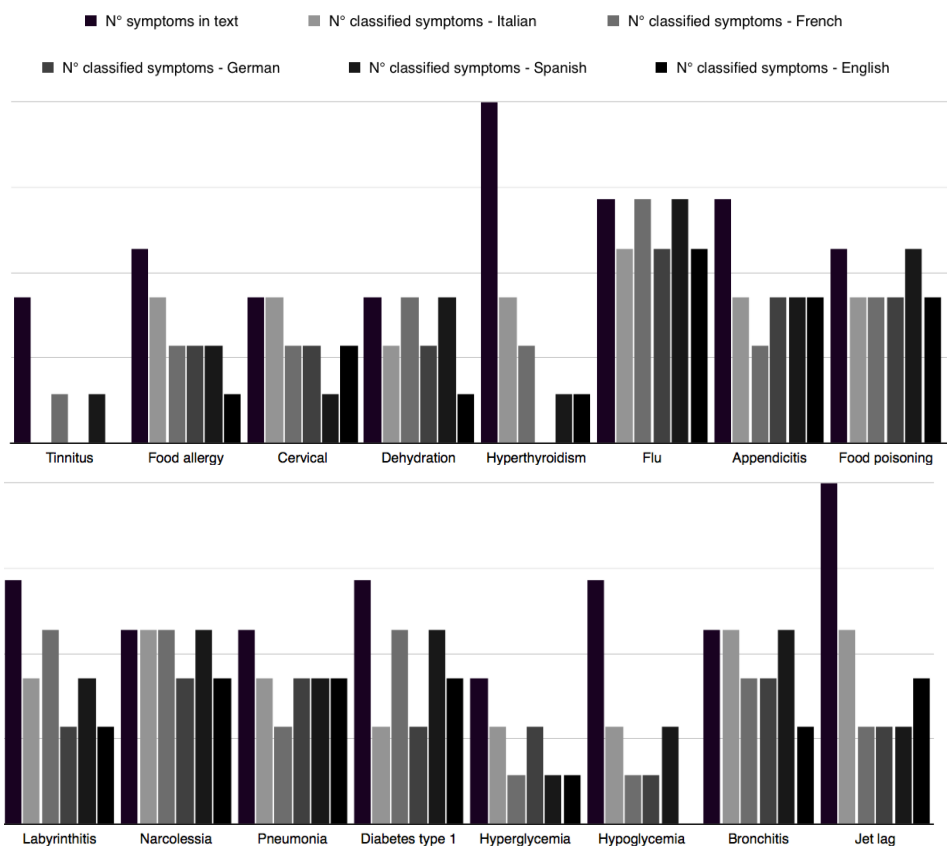

Fig. 3. For each disease, the leftmost column (in black) measures the average number of symptoms that should have been identified; the next five columns show the average number of correctly identified symptoms in Italian, French, German, Spanish and English sentences respectively.

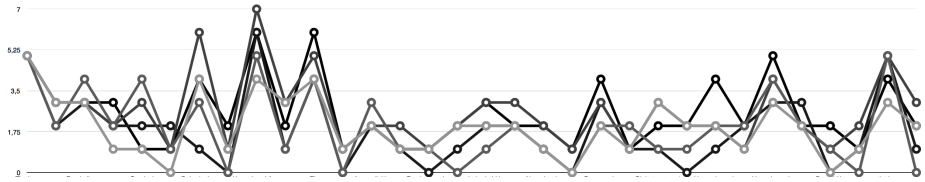

Fig. 4. Trend of errors for disease in the five languages (False Negatives). On the x axis the diseases are reported (labels are omitted) and on the y axis the number of false negatives for disease is reported: each line in the graphic is associated with one language.

language), the identification fails. This problem is due to the way the root of a word is computed, and to the way words are managed in BabelNet.

What emerges from Figure 4 is that false negatives have a very similar behavior despite the language of the sentence. This is again due to the two reasons discussed above. Despite these problems, which have a clearly understood motivation and which can be addressed by extending the ontology and by refining the management of word root extraction, MOoD-TC has demonstrated to be a flexible and ready-to-use approach for multilingual symptoms identification driven by a standard ontology we retrieved on the web.

6 Conclusions and Future Work

In this paper we presented the MOoD-TC architecture showing its possible use in the symptoms identification problem. The speech-to-text pre-processing stage can be faced using existing tools, and the post-processing stage with a translation of the identified symptoms into the doctor's language can be addressed using BabelNet, in the same way we exploit BabelNet for bridging the text, whatever its language, and the ontology. The more challenging post-processing stage of supporting the user in providing a diagnosis given a set of identified symptoms could be addressed by means of sophisticated expert system such as the old and well known MYCIN [4] and more recent projects (http://www.easydiagnosis.com/, https://www.diagnose-me.com/, [2]), some of which are ontology-driven [1].

Our framework does not face many well known open problems in multilingual text classification and information extraction such as negation [23] and named entities, but rather it provides a flexible and modular approach ready for integrating, with limited effort, the results and algorithms addressing the above problems coming from the research community.

In the short time, our work will be devoted to overcome the problems that limit the performances of MOoD-TC in the considered scenario: we will make the word identification more flexible and we will extend the symptoms ontology with those symptoms which have not been modeled so far.

In the future, it would be interesting to run an experimental comparison between our approach and a machine learning one. In case of a limited number of labeled examples, in fact, it would be feasible to apply semi-supervised learning methods. Depending on the comparison results, we will also consider to combine both approaches, using a domain ontology to improve the results of a traditional machine learning approach.

References

1. B. Al-Hamadani. CardioOWL: An ontology-driven expert system for diagnosing coronary artery diseases. In *2014 IEEE Conference on Open Systems (ICOS)*, pages 128–132, 2014.
2. R. P. Ambilwade, R. R. Manza, and B. P. Gaikwad. Medical expert systems for diabetes diagnosis: A survey. *Int. J. of ARCSSE*, 4(11), 2014.
3. S. Beux. MOoD-TC: A general purpose multilingual ontology driven text classifier. Master's Degree Thesis in Computer Science, University of Genova, Italy, 2015.
4. B. G. Buchanan and E. H. Shortliffe. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, 1984.

5. G. de Melo and S. Siersdorfer. Multilingual text classification using ontologies. In *ECIR Conference, Proceedings*, volume 4425 of *LNCS*, pages 541–548. Springer, 2007.
6. D. Demner-Fushman, W. Chapman, and C. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009.
7. C. Doulaverakis, G. Nikolaidis, A. Kleontas, and I. Kompatsiaris. Panacea, a semantic-enabled drug recommendations discovery framework. *J. Biomedical Semantics*, 5:13, 2014.
8. Global Reach. Global internet statistics (by language). Technical report, Global Reach, June 2005.
9. H. W. Griffith. *Complete guide to symptoms, illness & surgery for people over 50*. Body Press/Perigee New York, NY, 1992.
10. B. Guo-Wei and C. Hsin-Hsi. Cross-language information access to multilingual collections on the Intenet. *Journal of the American Society for Information Science*, 51, 2000.
11. H. Harkema, I. Roberts, R. Gaizauskas, and M. Hepple. Information extraction from clinical records. In S. Cox, editor, *Proceedings of the 4th UK e-Science All Hands Meeting*, Nottingham, UK, 2005.
12. P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization*. John Benjamins, 2002.
13. M. Leotta, S. Beux, V. Mascardi, and D. Briola. My MOoD, a multimedia and multilingual ontology driven MAS: design and first experiments in the sentiment analysis domain. In *ESSEM Workshop, Proceedings*, pages 51–66, 2015.
14. S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–144, 2008.
15. P. Nadkarni, L. Ohno-Machado, and W. Chapman. Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.
16. R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.
17. D. W. Oard and B. J. Dorr. A survey of multilingual text retrieval. Technical report, College Park, MD, USA, 1996.
18. A. Rosier, A. Burgun, and P. Mabo. Using regular expressions to extract information on pacemaker implantation procedures from clinical reports. In *Proceedings of the AMIA Annual Symposium*, Washington DC, USA, 2008.
19. B. R. South, S. Shen, M. Jones, J. H. Garvin, M. H. Samore, W. W. Chapman, and A. V. Gundlapalli. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics*, 10(S-9):12, 2009.
20. A. Stubbs, C. Kotfila, H. Xu, and Özlem Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics*, 58, Supplement:S67 – S77, 2015.
21. O. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5):552–556, 2011.
22. K. B. Wagholikar, M. Torii, S. Jonnalagadda, H. Liu, et al. Pooling annotated corpora for clinical concept extraction. *J. Biomedical Semantics*, 4:3, 2013.
23. M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 60–68, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.