UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Semantic annotation of electronic health records in a multilingual environment

Luís Campos

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIDADE EM BIOINFORMÁTICA

2017

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Semantic annotation of electronic health records in a multilingual environment

Luís Campos

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIDADE EM BIOINFORMÁTICA

Tese orientada pelo Prof. Doutor Francisco Couto e pelo Dr. Vasco Pedro

2017

Resumo

Palavras Chave: palavras chave

Abstract

Keywords: keywords

Resumo Alargado

Acknowledgements

Contents

1 Introduction **1**

1.1 Motivation . . . 1
1.2 Objectives . . . 2
1.3 Methodology . . . 3
1.4 Contributions . . . 4

2 Related Work **5**

2.1 Text Mining . . . 5
2.1.1 Named-entity Recognition . . . 5
2.1.2 Application of Text Mining on Radiology Reports . . . 6
2.1.3 Ontologies . . . 7
2.2 Translation . . . 8
2.2.1 Machine Translation . . . 8
2.2.1.1 Word-Based Models . . . 8
2.2.1.2 Phrase-Based Models . . . 10
2.2.2 Post-editing . . . 11
2.2.3 Machine Translation Services . . . 11
2.2.3.1 Yandex . . . 11
2.2.3.2 Google . . . 12
2.2.3.3 Unbabel . . . 12
2.2.4 Translation of Medical Text . . . 12
2.2.4.1 Multilingual Text Mining . . . 12
2.2.4.2 Machine Translation of Doctor-Patient Communication . . . 13
2.2.4.3 Machine Translation of Public-Health Information 14

2.2.4.4 Machine Translation for Information Retrieval . . 14
2.2.4.5 Machine Translation of Other Types of Medical Text 14
2.2.5 Translation of Ontologies 15
2.3 External Tools and Terminologies 16
2.3.1 RadLex 16
2.3.2 Open Biomedical Annotator 18
2.3.3 NOBLE Coder 18
2.4 Evaluation Metrics 20
2.4.1 Micro- and Macro- Evaluation Metrics 22

3 Framework **23**
3.1 Portuguese-English Parallel Corpus 23
3.1.1 Web Crawl of the articles (1,2) 23
3.1.2 Note On Human Translations 25
3.1.3 Yandex Translation (3) 25
3.1.4 Google and Unbabel Translation (4,5) 25
3.2 Annotation 25
3.2.1 Direct Match - Annotation with NCBO Annotator 26
3.2.2 All Match and Best Match - Annotation with NOBLE Coder 26
3.3 Evaluation 27

4 Experimental Results **29**
4.1 NER Gazetteer-based approach 29
4.1.1 Clinical Finding Subset 32
4.1.2 Anatomical Entity Subset 32
4.1.3 Imaging Modality Subset 33
4.1.4 Imaging Observation Subset 33
4.1.5 Non-Anatomical Substance Subset 33
4.1.6 Object Subset 33
4.1.7 Procedure Subset 34
4.1.8 Procedure Step 34
4.2 Discussion 34
4.3 Conclusions 34

5 Conclusions **35**

References **37**

List of Figures

List of Tables

2.1 Lexical translation probability table for the word broken 9
2.2 NOBLE matching strategies present in the GUI interface. Adapted from (Tseytlin *et al.*, 2016). This correspond to the options used in the GUI tool. 20

3.1 Number of articles by journal in parallel corpus 24

4.1 Number of RadLex terms found by document 29

Chapter 1

Introduction

1.1 Motivation

Radiology reports describe the results of radiography procedures (e.g., X-ray imaging) and have the potential of being an useful source of information, which can bring benefits to health care systems around the world. But because these reports are written in a free-text mode, it is hard to extract information automatically from them. This is more problematic when these reports were mostly stored in physical format (paper) - the fact that these reports can now be accessed digitally make them amenable for processing using Text Mining techniques.

A lot of work has been done on this research area (Pons *et al.*, 2016), but it is usually assumed that the reports are written in English. For example, (Hassanpour & Langlotz, 2016) created an information extraction system for English reports that depends on RadLex, a lexicon for radiography terminology, which is freely available in English. Because of this, the system can not be applied to reports written in other languages. And even if the system was not dependable on an English lexicon, it is not certain if the results would be the same if text in another language was used.

Assuming that the information automatically extracted from radiology reports using Text Mining techniques can bring benefits to health care systems, this waste of information caused by language barriers can potentially have a negative impact on everyone's health.

If what is wanted is to use tools that depend on English lexicons, one possible solution to the problem could be to translate the lexicon itself (Bretschneider *et al.*, 2014). Other obvious solution, and the one explored in this thesis, is to translate the reports. If the translation is done by professionals trained in the translation of medical texts, we probably can assume that not much information is lost in translation. We call this type of translation Human Translation (HT). But professional translators are expensive and there are a lot of reports to translate, so this would have a really high monetary cost. Another option is to use Machine Translation (MT). Notwithstanding the lower translation quality, it is way cheaper and more feasible in a large scale. Finally, an option that tries to get the best of both worlds is using Machine Translation with Post-Editing (MT-PE) by humans. Basically, the text is automatically translated by a machine and then the translation is corrected by a human. Cheaper than the HT option and with better quality than the MT one.

So, how much information is modified or lost during MT or MT+PE compared to HT, affecting the results of Text Mining tools? To the best of my knowledge, the most similar work to this one is (Castilla, 2007). He founds that a rule-based MT system has a good performance in translating Portuguese text to English for the purposes of applying a text mining tool (better described in 2.2.4.1). The author doesn't compare translation systems, something that is done on the present work.

1.2 Objectives

In this thesis I have studied how MT and MT+PE compares with HT on the simple task of Named-entity recognition (NER) using a gazetteer-based approach, this gazetteer consisting of RadLex terms. The identification of RadLex terms can be useful, for example, in image retrieval (Gerstmair *et al.*, 2012) systems, so this is not just a toy example.

If I have non-English radiology reports and I want to translate them so that I can identify RadLex terms for use on some other system, what kind of translation should I use? In this thesis, I try to help to answer this question.

Hypothesis: MT+PE is a good trade-off between quality and cost, compared with MT and HT, for translating radiology reports for the purpose of identifying RadLex terms.

For this to be true, these conditions have to hold:

1. MT+PE has to be cheaper than HT
2. MT+PE quality for the task at hand has to be close enough to HT quality
3. MT+PE quality for the task at hand has to be better than MT quality, enough to compensate its higher cost

The last condition its important because if MT+PE quality is similar to MT quality, as MT cost is lower, maybe it's worth to just use MT. In this thesis I only try to answer to the quality issues, not doing a thorough economic analysis of the problem.

1.3 Methodology

To answer these questions I've compared the results of the NER task on the MT and MT+PE translations with the results of the NER task on the HT translation. Assuming that the HT translation is the right translation, the closer the results are to the ones of the HT, the better the translation for the task at hand. So, for the proposed hypothesis to be true, the results of the NER task on the MT+PE task have to be closer to the results of the HT translation than the results of the MT translations.

For this purposes I've created a parallel corpus containing a number of scientific articles related to radiology and corresponding HT, MT and MT+PE translations. These reports were annotated with RadLex terms using NOBLE Coder and a custom built tool. The annotations of the MT and MT+PE translations were then compared with the ones from HT.

1.4 Contributions

Thus, the following specific contributions can be enumerated as follows:

Contribuition1:

Chapter 2

Related Work

2.1 Text Mining

Text Mining consists in the machine supported analysis of text (Hotho *et al.*, 2005). It can be used, for example, to help researchers cope with information overload (Cohen & Hersh, 2005) due to the big volume of scientific data in the form of unstructured literature. More related to this thesis, it can also be used to extract information from free-text radiology reports (Pons *et al.*, 2016).

Because Text Mining has to manipulate text, it is not too surprising that it borrows tools from Natural Language Processing (NLP), a research fields that seeks to improve computational understanding of natural language.

In the next subsections I will explain one of these NLP tools, called Name-entity recognition and briefly explore how Text Mining can be used to extract information from Radiology Reports.

2.1.1 Named-entity Recognition

Named-entity recognition (NER) is a task of NLP that has the goal of locate and classify all the named-entities in a certain document. Named-entities are elements of the text that belong to one of certain predefined classes. For example, in the phrase *Atrial fibrillation has strong associations with other cardiovascular diseases* the term *Atrial fibrillation* is a named-entity that belongs to the class *Disease*.

The task of NER can be further divided in two subtasks: identify where in the text are the named-entities and classify the named-entities.

The approaches of NER can be divided into three categories (Mansouri *et al.*, 2008): Rule-based approaches, Machine Learning based approaches and hybrid approaches.

- In rule-based approaches the identification and classification subtasks are based on rules crafted by humans. Usually domain specific.
- In ML based approaches the subtasks are turned into classification problems and machine learning algorithms are used to identify and classify named-entities. These approaches are easily ported to different domains other than the ones they were originally developed to be applied on.
- Hybrids approaches combines the two last approaches.

Gazetteer based-approaches are a subset of the rule-based approaches. In this approach we already have a list of the named-entities (a gazetteer) that we want to identify in the text. For example, if we want to identify chemical entities in text, we use a chemical-entities gazetteer. The goal of the gazetteer based-approach is then to identify, in text, mentions of terms presented in the gazetteer. This could be done by direct matching, as implemented by the Open Biomedical Annotator [1] (Jonquet *et al.*, 2009). In this strategy, the system only tries to find in text terms that are also in the gazetteer, not considering, for example, lexical variations. The recall can be lower than expected because lexical variants (like plurals), abbreviations and partial matchings of gazetteer terms are not recognized in the text. For this purpose, more complex tools like NOBLE Coder[2] (Tseytlin *et al.*, 2016) or Concept Mapper (Stewart *et al.*, 2012) can be used.

2.1.2 Application of Text Mining on Radiology Reports

Text Mining tools can be used for automatic detection of important findings in Radiology Reports. For example, (Dreyer *et al.*, 2005) used an algorithm based

[1] `http://bioportal.bioontology.org/annotator`
[2] `http://noble-tools.dbmi.pitt.edu/`

on information theory to classify reports as having/not having important clinical findings and as having/not having recommendations for subsequent action. (Cotik *et al.*, 2015) did something similar for Spanish reports, using a translation of RadLex terms. These tools can also be used to detect the presence of more specific findings, as the presence of invasive mold diseases (Ananda-Rajah *et al.*, 2014) or invasive fungal diseases (Martinez *et al.*, 2015), both using a classifier based on a Support Vector Machine. Also possible is to extract general information about reports (Hassanpour & Langlotz, 2016) and the data obtained can be used as input to other tools.

In literature it's possible to find some examples of Radiology reports/images search applications, that use NLP tools. The goals of these search tools include search for educational, research and clinical decision support purposes. One example of such a system is Render (Dang *et al.*, 2009), which even applies one of the information extraction system mentioned above (Dreyer *et al.*, 2005) to improve relevance of information retrieved.

Other applications include studying the appropriateness of existing Radiology reports templates, as done by (Hong & Kahn, 2013)

2.1.3 Ontologies

To answer the questions presented in Chapter 1, the RadLex ontology is used. In this section I briefly explain what is an ontology. An ontology is a "common, controlled knowledge representation designed to help knowledge sharing and computer reasoning" (Robinson & Bauer, 2011). It is a way to represent a subset of the real word which can be used as basis for communication between parties wanting to change information about that subset of the real word.

RadLex, for example, is a representation of the subset of the world related to radiology which can be used as standard on how to talk about radiology. Ontologies usually have a tree structure in which a class, representing some abstract entity in the real world, can have subclasses. For example, in RadLex, there is the class *clinical finding* which has subclasses *benign finding* and *pathophysiologic finding* (among others). This subclasses have a *is a* relationship with their parent

classes: *benign finding* is a *clinical finding*. Other common relationship used in ontologies is the *part of* relationship.

Other popular examples of ontologies include the Gene Ontology[1], focused on genomics, SNOMED CT[2], a healthcare related ontology and ChEBI[3], an ontology of small molecular entities.

2.2 Translation

2.2.1 Machine Translation

Machine Translation (MT) is the use of computers to automatically translate natural language text. Currently, Statistical Machine Translation (SMT) is the most popular approach to MT. Other approaches included Rule-Based Machine Translation (RBMT) and Neural Machine Translation (NMT). RBMT involves the use of hand-crafted rules on how to do the automatic translation and NMT uses neural-networks and it's use has recently been growing (Bentivogli *et al.*, 2016). I will now briefly review word-based and phrase-based which are both covered by the SMT approach. This is mostly based on (Koehn & Philipp, 2010).

2.2.1.1 Word-Based Models

These kind of models are not the state of the art anymore, but many of the principles and techniques of this approach are still in use today. The idea here is to translate the sentences word by word. Here is an example, translating English to Portuguese:

English - The bone is broken
Portuguese - O osso está partido

This is easy for a human to translate, but how would a computer know that *partido* is the translation of *broken* when *broken* has other potential translations? For example, the word *broken* could be interpreted as being financially ruined, as

[1] http://www.geneontology.org/
[2] http://www.snomed.org/snomed-ct
[3] https://www.ebi.ac.uk/chebi/

in “I’ve spent all the money in the casino, I’m completely *broken*". In that case, *broken* would be translated to *falido*. Of course, this doesn’t make sense because bones don’t have a financial life but the computer doesn’t know that.

One way to teach the computer which translation to use would be to pick a large collection of English texts paired with the corresponding Portuguese translation and check how many times *broken* is translated to *partido* and how many times it is translated to *falido*. Lets assume that in our collection of texts the word *broken* is translated to *partido* 80% of the times and to *falido* 20% of the time. With this we could create a lexical translation probability table for the word *broken*. We could have a table like this one for every word in the source texts.

broken	
t	p(t|s)
partido	0.8
falido	0.2

Table 2.1: Lexical translation probability table for the word broken

Here *t* stands for target, *s* stands for source and *p(t|s)* is the probability that the target word is the translation of the source word. So, when the computer is translating the sentence above and arrives to the word *broken*, it checks the table and chooses *partido* as the translation because it has the higher probability of being the real translation. This type of estimation is called maximum likelihood estimation. What we are doing here is estimating lexical translation probability distributions.

The example above was easy because the sentences were aligned word by word. This is not always the case. For example, the English expression *red swelling* should be translated to *inchaço vermelho*, not *vermelho inchaço*[1]. Meaning, sometimes we must do some word reordering so that the translation is correct. This is accommodated by using an alignment model. But how can we generate an alignment model from a pair of collection of texts if we don’t know which word is aligned with which word? This is done by using the expectation maximization algorithm, which, in this case, iteratively applies the alignment model to the texts

[1] red -> vermelho, swelling -> inchaço

(expectation step) and learns the alignment model from the texts (maximization step) until convergence of the parameters in the algorithm.

With the lexical translation probability distributions and an alignment model we have a translation model. But this is not enough. A translation could be syntactically and semantically right but still not sounding right. For example, two possible translations of *chá forte* are *strong tea* and *powerful tea*. However, the second option doesn't sound right, it is not fluent. This problem is solved by using a language model. With an English language model, for example, we could calculate the probability that a certain sentence is correct English, considering all the data that was used to train the model. A language model would probably give a low probability to the phrase *powerful tea* because normally the word powerful is not used with the word *tea*.

We combine the language model and the translation model this way:

$$\arg\max_t \Pr(t|s) = \arg\max_t \Pr(s|t)\Pr(t) \tag{2.1}$$

We want to find the target word (t) with the higher probability of being the translation of the source word (s). $\Pr(t|s)$ represents the translation model and the the $\Pr(t)$ represents the language model. This way of combining the translation and the language models is called noisy-channel model.

2.2.1.2 Phrase-Based Models

In this approach, instead of translating a sentence word by word we translate small words sequences at a time, sequences that we call phrases. These models have a better performance than the word-based models and this is not too surprising. Sometimes words are not the best unit of translation: there are cases when two words in the source sentence are translated into one word in the target sentence, for example. Another advantage is that translating phrases instead of words can help to solve ambiguities, as in the problem of deciding how to translate the text *chá forte* (see last section). We would check a parallel collection of texts and realize that most of the times *chá forte* is translated to *strong tea*. So, the idea here is to divide the sentence in phrases, translate the phrases and do some reordering if necessary.

2.2.2 Post-editing

Post-editing (PE) is the task of editing, modifying and/or correcting a text that was pre-translated by use of MT, in order to improve the translation. (Somers, 2003) refers to the lower cost of MT+PE compared with HT to explain the growth of PE: companies want to become global but can't afford the cost of HT to translate from native language to the many languages they want to operate on.

(Koponen, 2016) tried to understand if MT+PE is really worth, compared with just HT, concluding that yes, most of the times it is worth it, but it depends on the quality of the MT, which in turn depends on, for example, the quality of the MT system and on the language pair.

Most of the research regarding PE refers to work done by professional translators. One approach that has been gaining traction is the use of the crowd to do the PE (Tatsumi & Aikawa, 2012). The advantages of this strategy include lower per-word cost and sometimes an higher speed. One big disadvantage is less assurance of quality.

2.2.3 Machine Translation Services

2.2.3.1 Yandex

Yandex[1] is a Russian search-engine company. At the time the work for this thesis was being done, Yandex.Translate (the name given to Yandex's MT system) uses a statistical approach. From their website[2], the system is composed by three components, a translation model, a language model and a decoder which is the part that actually does the translation.

I couldn't find any research paper evaluating the translation's quality of Yandex.Translate in the language pair Portuguese-English.

[1] https://yandex.com/

[2] https://tech.yandex.com/translate/doc/intro/concepts/how-works-machine-translation-docpage/

2.2.3.2 Google

Google[1] is a company from the United States that offers a lot of technological services, including machine translation. For the language pair Portuguese-English, their translation services now use Neural Machine Translation[2] (see section 2.3.1).

2.2.3.3 Unbabel

Unbabel[3] is a Portuguese start-up which offers translation services focused on conversational content like costumer service or websites copywriting, using an MT+PE approach. Although it's not mentioned in the Unbabel's API documentation, for the language pair Portuguese-English, currently Unbabel uses Google Translate's services in MT step of the MT+PE approach (personal communication). Next is an overview of Unbabel's translation pipeline:

1. Text is translated by MT (in this case, using Google Translate)
2. MT translated text is post-edited by users of the Unbabel platform. Users translate the text using Unbabel's web-interface or mobile app.
3. Translation resulting from last step is reviewed by an Unbabel's senior user, an user that was promoted for having good ratings

From now on I'm going to call this type of translation *Unbabel Translation*.

2.2.4 Translation of Medical Text

2.2.4.1 Multilingual Text Mining

There is not much research studying the effect of translation on NLP techniques. (Castilla, 2007) is the most similar work to the one developed on this thesis and curiously, also studies translation of Portuguese medical text. In the main part of the study, Portuguese-written radiology reports were translated to English using the SYSTRAN MT system, which uses a rule-based approach, complemented

[1] https://www.google.com
[2] https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-t
[3] https://unbabel.com/

with a specialized medical translation dictionary and then the translation was processed by the Medical Language Extraction and Encoding System (MEDLEE) to extract information on the presence of mentions of certain medical conditions. The results were compared to reference results created by three radiologists on the original reports. The results are really positive, with values of sensitivity, specificity, positive and negative predictive values all above 88%. These results suggest that for this specific task of information extraction a MT translation retains a lot of information from the original text.

2.2.4.2 Machine Translation of Doctor-Patient Communication

Most of the work done on medical translation focuses on translation of doctor-patient communication. This has the objective of breaking language barriers that sometimes exist between a doctor and a patient who don't speak the same language, with health-related consequences to the patient (Schyve, 2007). This could be done with trained medical interpreters but that option is costly compared with using MT and raises problems regarding patient confidentially.

Several MT speech-to-speech translation systems for doctor-patient communication exist, but for most of them, evaluations are not found in the literature. One exception is (Bouillon *et al.*, 2005) which studies MedSLT, a multilingual spoken language translation system tailored for headache, chest pain and abdominal pain domains. However, (Bouillon *et al.*, 2005) only studies the appropriateness of the design choices within the system, not comparing its performance with anything else. Others example of systems of this type are Jibbigo[1], Universal Doctor[2] and Transonics (Nagata & Pedersen, 2005).

(Kaliyadan & Pillai, 2010) did a small study on the use of Google Translate to translate between English and French during doctor-patient interaction in India medical offices, with promising results regarding patient satisfaction. Also using Google Translate, (Patil & Davies, 2014) studied the quality of the translation of 10 commonly used medical statements to 26 languages. Of all the 260 translations, 57.7% were right. The results were better for Western European languages than for others. Portuguese had the highest score, with 9 of the 10 sentence

[1] http://jibbigo-translator-2-0.soft112.com/
[2] http://www.universaldoctor.com/

translated being right. Other work was also done on non-European languages, which have less resources (Kathol *et al.*, 2005; Musleh *et al.*, 2016).

Some researchers (G *et al.*, 2013; Kaliyadan & Pillai, 2010; Marta R. Costa-jussà, Mireia Farrús, 2012) suggest that MT should be used very cautiously in this situations, because of imperfect performance in a domain where accuracy is really important. One way to improve the systems could involve the use of existing public medical terms database (Eck *et al.*, 2004).

2.2.4.3 Machine Translation of Public-Health Information

In the USA, most of the public health information is written in English, although a substantial percentage of the population have limited English proficiency. One of the barriers for more widespread translation is the cost of translation services and a way of streamlining the process would be using MT+PE. (Kirchhoff *et al.*, 2011; Turner *et al.*, 2015) studied the feasibility of this system for translation from English to Spanish, with some promising results, and to Chinese, which was more problematic.

2.2.4.4 Machine Translation for Information Retrieval

The ACL 2014 Ninth Workshop on Statistical Machine Translation had a Medical Translation Task (Bojar *et al.*, 2014), which consisted in two subtasks: translation of sentences from summaries of medical articles and translation of queries entered by users of medical information search engines. This task was supported by the Khresmoi [1] project which develops a multilingual search and access system for biomedical information and documents, allowing the user to make search queries and read summaries of the results in their own language. The task had 8 participants, the winner being the UEDIN team (Durrani *et al.*, 2014) which used the Moses phrase-based system.

2.2.4.5 Machine Translation of Other Types of Medical Text

Studies of the translation of other types of documents are also present in the literature. For example, (Wołk & Marasek, 2015) compares neural based with

[1] `http://khresmoi.eu/`

statistical machine translation of descriptions of medical products in the language pair Polish-English, obtaining mixed results.

More related to the work done on this thesis, (Castilla, 2007) studied the use of the MT application SYSTRAN to translate sentences from radiology reports. The MT system uses a ruled-based approach and was complemented with a specialized medical translation dictionary. The translations were evaluated by an expert in the field, finding good scores for understandability, fidelity with original text and translation coverage of the original text.

(Zeng-Treitler *et al.*, 2010) tested if a general-purpose machine translation tool like the Babel Fish is adequate to translate sentences of discharge summaries, surgical notes, admission notes, and radiology reports from English to Spanish, Chinese, Russian and Korean. They found that most of the times the translation is incomprehensible and inaccurate.

More recently, there was a Biomedical Translation Task during the ACL 2016 First Conference on Machine Translation (WMT16) in which the participants were asked to submit systems to translate titles and abstracts from scientific publications (Bojar *et al.*, 2016). The evaluators note that the quality of the machine translation is still poor in comparison to the reference translations. The only submission to the English-Portuguese and Portuguese-English translation tasks (Aires *et al.*, 2016) were the ones with the worse results relative to the baseline system.

2.2.5 Translation of Ontologies

One alternative solution to the one I'm exploring in this thesis, translating the medical text to English, is to translate the lexicon, on which the task at hand depends on, to the language of the medical documents we want to study. For example, if a researcher have a non-English corpus and wants to annotate it with terms of some ontology, it will be a problematic task since most of the available ontologies are not multilingual. To solve this the researcher could translate the ontologies she wants to use to the language of the corpus. There is some work on this problem of translating ontologies, but only one article was found related to

biomedical ontologies (Bretschneider *et al.*, 2014). Having in mind that translating all the entries of an ontology one wants to use would be expensive, the authors propose translating only a subset of the ontology, a subset relevant to the task at hand and doing this semi-automatically with the help of the corpus one wants to annotate. With this, the authors improved the annotation of German text with RadLex terms.

2.3 External Tools and Terminologies

Some of the work done during the thesis used and was inspired by some external tools and terminologies that I now briefly review.

2.3.1 RadLex

RadLex[1] is a domain-ontology which focuses on radiology-related terms. It was developed to standardize annotation, indexation, and retrieval of radiology information resources in the digital world (Langlotz, 2006) and it helped to fill a gap in radiology terminology (Langlotz & Caldwell, 2002; Woods & Eng, 2013). The RadLex terms were originally gathered from existing ontologies at the time, including the American College of Radiology (ACR) Index, SNOMED-CT, and the Foundational Model Anatomy and it is a highly dynamic ontology: its number of terms grew from around 8000 to around 75000 in just ten years.

There are a few studies on the completeness of RadLex. (Marwede *et al.*, 2008) found that an old version of RadLex covered 84% of terms extracted manually from 250 thoracic CT reports, with higher coverage for terms in the *Findings* (90%) category and lower coverage for the *Modifier* category (78%). Curiously, in a study using more recent versions of RadLex (versions 3.1–3.5) (Woods & Eng, 2013) found a lower coverage of 62% using the same type of reports (they used less reports in this study, just 100). They find higher coverage for the categories of *anatomic objects* and *physiological conditions* and lower coverage for the categories of *imaging observations* and *procedures* (the categories used in both studies are not the same). The authors justify the lower coverage with the

[1] `http://www.rsna.org/RadLex.aspx`

inclusion in the study of categories such as *procedures*, which didn't had any match with RadLex terms. They also used a different methodology to find matches between manual extracted terms and Radlex terms. These studies analyzed the coverage of RadLex of terms mentioned in the contents of radiology reports. (Hong *et al.*, 2012), on the other hand, studied how well RadLex covers the terms of templates of structured radiology reports developed by the Radiological Society of North America, finding that 41% of the terms found in the templates matched exactly to RadLex and that 26% matched partially. Since these analysis, new versions of RadLex were launched so the results and critics present in the studies could not be that relevant anymore.

One could use RadLex to assist in the matching of articles manuscripts to reviewers profiles, like done by the RadioGraphics journal (Klein, 2013). Or to help in the visual analysis of neurography images (Wang *et al.*, 2015). Having said this, most of the examples described in the literature are of applications related to Information Retrieval (IR), the task of extracting some information resource from a collection of information resources. These resources can be images or websites, for example. One such example of a IR system using Radlex, is (Spanier *et al.*, 2016), who takes advantage of the tree structure of this ontology to create a new method of case-based image retrieval (M-CBIR). Most existing M-CBIR systems use low-level characteristics of medical images (like color, shape and texture) to induce similarity between them. But this is problematic since medical images which show the same type of content can have different low-level characteristics. One solution is to induce this similarity from the information contained in the textual radiological reports that accompany the images and the authors take advantage of RadLex to do just that. This can help radiologists to find related medical cases in a certain database which then can help them in their decision-making process. Other approaches to IR systems using Radlex include the ones described in (Do *et al.*, 2010), (Kurtz *et al.*, 2014) and (Gerstmair *et al.*, 2012).

2.3.2 Open Biomedical Annotator

The Open Biomedical Annotator (OBA)[1] is an open-source tool for NER using a gazetteer-based approach, made available by the North-American National Center for Biomedical Ontology (Jonquet *et al.*, 2009), which can be used to annotate text with concepts from ontologies. For example, if you go to the website, input a radiology report and choose the ontology RadLex, the tool will return all the mentions in the text of terms belonging to the RadLex terminology. It can easily be used as a web-service and it is relatively fast. It uses a case-insensitive direct match approach, not considering lexical variations of words (see 2.1.1).

2.3.3 NOBLE Coder

NOBLE Coder[2] (Tseytlin *et al.*, 2016) is a software for NER using a gazetteer-based approach. The gazetteer is set by the user (it has to be in UMLS (RRF)[3], OWL[4] or OBO[5] formats or be present in BioPortal[6]) and NOBLE finds, in an arbitrary text, mentions of terms found in the gazetteer.

Unlike the system used by OBA, NOBLE can find mentions of lexical variations of the terms present in the gazetteer because it applies word stemming. For example, *lobe* is a term present in the RadLex terminology, but it's plural, *lobes*, isn't. However, NOBLE considers that *lobes* is a mention of the term *lobe*, which is right. But this can sometimes go wrong; for example, NOBLE considers that *headings* is a mention of the RadLex term *head*, which is wrong. So although this strategy can improve recall it does so at the cost of precision.

The NOBLE tool is flexible in what is considered a mention of a gazetteer term, giving the user the power to adapt the tool for her specific purposes. This can be done by choosing to use or not a certain *matching option*. These include:

- **Subsumption** - Only matches the longest mention. For example, *toe*, *toe skin* and *skin* are all RadLex terms. If the "Subsumption" option is set, in

[1] http://bioportal.bioontology.org/annotator
[2] http://noble-tools.dbmi.pitt.edu/
[3] https://www.ncbi.nlm.nih.gov/books/NBK9685/
[4] https://www.w3.org/OWL/
[5] http://www.geneontology.org/faq/what-obo-file-format
[6] http://bioportal.bioontology.org/ontologies

the text *toe skin*, only the term *toe skin* will be recognized. Otherwise, the terms *toe* and *skin* are also recognized.

- **Overlap** - ...
- **Contiguity** - Terms must be contiguous to be matched. For example, if set, in the text *multiple ducts lesions* both *multiple ducts* and *multiple lesions* are considered matches, although *multiple* and *lesions* are not adjacent to each other. It's possible to set how many irrelevant words can be between words belonging to a term (in 2.2, this is called *gap*).
- **Order** - Terms must be in the same order as in the gazetteer to be considered mentions. If not set, *lesions multiple* is considered a mention of the Radlex term *multiple lesions*.
- **Partial** - Partial match with terms in gazetteer are considered a gazetteer term mention. If set, *multiple* is considered a mention of *multiple lesions*.

The user can also choose to, for example:

- Skip single letter words
- Skip stop words
- Use heuristics to filter out potential false positives
- When a term can be considered a mention of more than one concept in the gazetteer, select only the highest scoring one

Different combinations of these options are useful for different purposes. NOBLE already offers some built-in matching strategies, listed in 2.2.

The authors of the tool provide suggestion for what kind of task each strategy is more appropriate. For example, they suggest that the *best match* strategy is best for concept coding and information extraction and that the *all match* strategy is more suitable for information retrieval and text mining.

(Tseytlin *et al.*, 2016) compares the NOBLE tool with other gazetteer-based NER tools, finding that its performance in recognizing terms from gazetteers

	Combination of matching options				
Task	**Subsumption**	**Overlap**	**Contiguity**	**Order**	**Partial**
Best match	Yes	Yes	Yes (gap=1)	No	No
All match	No	Yes	No	No	No
Precise match	Yes	Yes	Yes (gap=0)	Yes	No
Sloppy/Partial match	No	Yes	No	No	Yes

Table 2.2: NOBLE matching strategies present in the GUI interface. Adapted from (Tseytlin *et al.*, 2016). This correspond to the options used in the GUI tool.

its comparable with other similar software like Concept Mapper (Stewart *et al.*, 2012) or cTAKES [1] [2], although it probably depends a lot on the corpus used.

One big advantage of NOBLE is its ease of use compared with other similar systems. Little or no programming skills are needed to use the software since it includes a GUI (Graphical User Interface) which allows an user to upload gazetteers in a number of formats and easily annotate texts.

2.4 Evaluation Metrics

For a certain task (for example, annotation of terms that represent diseases from a corpus) it is useful to have standard evaluation metrics so that we can compare many systems and know which one is the best. In information retrieval and information extraction systems precision (P), recall (R) and F-score (F) are the measures that are mostly used. For example they were the measures used in a competition which involved a task similar to the example I gave above (Elhadad *et al.*, 2015).

To use this measures we need to have a reference, a gold-standard, which we assume represents the perfect performance in a certain task, the ground truth. In the example of extraction of disorder mentions, it could be an annotation done by an human expert. To calculate this measures we also need the number of true

[1] `https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.0+-+Dictionary+Lookup`

[2] `https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.2+-+Fast+Dictionary+Lookup`

positives, true negatives, false positives and false negatives. I will illustrate each one of these with the example of the annotation of diseases mentions.

- True positive (TP) – The system being tested annotated a term also annotated in the reference.
- True negative (TN) – The system didn't annotate a term that is also not annotated in the reference.
- False positive (FP) – The system annotated a term that is not annotated in the reference.
- False negative (FN) – The system didn't annotate a term that is annotated in the reference.

Precision corresponds to the fraction of the terms annotated by the system that are also annotated in the reference.

$$P = \frac{TP}{TP + FP} \tag{2.2}$$

If of the 10 terms annotated by the system, only 6 are annotated by the reference, then the system has a precision of 0.6. If every term extracted by the system is also extracted by the reference, then the system has a precision of 1, the best score possible. But the system can have a score of 1 if only annotates one right term, even though there are a lot of other terms annotated in the reference. This system, although having a score of 1, would not be very useful. Recall is a measure that helps to solve this issue.

Recall calculates what fraction of all terms annotated in the reference are annotated by the system.

$$R = \frac{TP}{TP + FN} \tag{2.3}$$

If the system annotates 8 terms of the 10 that are annotated in the reference, then it has a recall of 0.8. If it annotates all of them, it has a recall of 1, the perfect score. But, as is the case with precision, this measure also has problems. If the system annotates all the terms in a corpus, it will have a perfect score in

the recall measure, because it is sure to have annotated all the terms annotated in the reference, although it also annotated a lot of wrong terms.

As you can see, both measures have problems when used in isolation. One way to combine them is by using the F-score measure, that corresponds to the harmonic mean of precision and recall.

$$F - score = 2 * \frac{P * R}{P * R} \tag{2.4}$$

2.4.1 Micro- and Macro- Evaluation Metrics

Now imagine that you want evaluate your system on more than one document. How do you aggregate the metrics explained above? You can sum the TP, FP and FN values of each document and then use the Precision, Recall and F-Score formulas exposed above. With this approach, you would calculate the Micro Precision, Micro Recall and Micro F-score.

Another approach is to calculate Precision, Recall and F-Score for each document and then average for all documents. This would give you the Macro Precision, Macro Recall and Macro F-score values.

Chapter 3

Framework

3.1 Portuguese-English Parallel Corpus

For the purpose of this work, I've created a Portuguese-English parallel corpus of research articles related to radiology. For each research article there is:

1. Original Portuguese text
2. Human Translated English text
3. Machine Translated English text (Yandex)
4. Machine Translated English text (Google)
5. Machine Translation + Post-Editing English text (Google + Unbabel)

In the next few lines I will explain how I've constructed the corpus.

3.1.1 Web Crawl of the articles (1,2)

First, I needed a list of articles related to radiography that were available both in English and in Portuguese. To get this list I've used the NCBO Entrez Programming Utilities (E-utilities)[1] to query the PubMed database with the search query "portuguese[Language] AND english[Language] AND radiography[MeSH Major

[1] `https://www.ncbi.nlm.nih.gov/books/NBK25501/`

Topic] AND hasabstract[text]" (search done on 11/12/2016). The last filter is used to avoid getting texts for which only the title is available.

Then I programmatically crawled each article PubMed page to get the URL where the full article could be found. Most of the articles were hosted in SciELO[1] so for the sake of consistency I've only included in the corpus articles hosted in there.

For the purposes of this work, it made sense to only include articles for which the original language is Portuguese, so I've also filtered the corpus by this parameter.

Finally, I've programmatically crawled the articles SciELO pages to get both language versions of articles text. I've extracted from the HTML everything from the abstract until, but not including, the references/bibliography.

Three of the article contained were about surveys, containing to much vocabulary about radiology. They were excluded from the corpus.

What is left is a parallel corpus of 53 articles, distributed by journal in the following way:

Table 3.1: Number of articles by journal in parallel corpus

Journal	Number Of Articles
Arquivos Brasileiros de Cardiologia	26
Jornal Brasileiro de Pneumologia	14
Revista do Colégio Brasileiro de Cirurgiões	4
Brazilian Journal of Otorhinolaryngology	2
Arquivos Brasileiros de Cirurgia Digestiva	2
Revista Brasileira de Cirurgia Cardiovascular	2
Jornal da Sociedade Brasileira de Fonoaudiologia	1
Einstein (São Paulo)	1
Revista Brasileira de Reumatologia	1

The corpus has a total of 170137 words[2] the longer article having 12451 and the smaller 848. The articles have an average of 3210 words each.

[1] `http://www.scielo.br/`

[2] Tokenization done by NLTK's word_tokenize function (`http://www.nltk.org/`)

3.1.2 Note On Human Translations

It is not known for sure how exactly the original human translations were done, since some of the articles are not recent and some of the journals did not answer my emails questioning about this, but all the answers received mentioned the use of specialized translation services. Having said this, we assume that the translations are of high quality since they were published by scientific magazines.

3.1.3 Yandex Translation (3)

The Portuguese version of the articles were machine translated using Yandex's free Translate API[1]. Each translation request had a limit of 10000 characters so an algorithm was used to break the text to various pieces, without breaking the text in the middle of sentences, send the translation request for each piece and then join everything back.

3.1.4 Google and Unbabel Translation (4,5)

Both MT with Google and MT+PE with Unbabel were obtained using Unbabel's API[2]. The requests for Unbabel Translations have a limit of words, so an algorithm similar used for the Yandex Translations was used.

3.2 Annotation

All the English versions of the articles in the corpus were annotated thrice with RadLex terms, one time using a direct matching approach and two using two of the built-in matching strategies provided by NOBLE Coder. I'm calling the three approaches Direct Match[3], All Match and Best Match[4]. Three different kinds of approaches were used to check what effect the annotation strategy have on the results.

[1] `https://tech.yandex.com/translate/`
[2] `http://developers.unbabel.com/`
[3] See 2.1.1
[4] See 2.3.3

Each class of the RadLex ontology has a *preferred name* and a list of synonyms. For all the approaches the output of each annotation process consists in the set of the preferred names of the RadLex terms that are mentioned in the corresponding article. I normalize all the mentions to the preferred name so that a use of the preferred name in one translation and the use of one of the synonyms in another translation are considered mentions of the same term.

3.2.1 Direct Match - Annotation with NCBO Annotator

The articles were annotated with the NCBO Annotator using the REST API[1]. The default parameters were used, namely:

- **expand_class_hierarchy** - false
- **expand_mappings** - false
- **minimum_match_length** - 3
- **exclude_numbers** - false
- **whole_word_only** - true
- **exclude_synonyms** - false
- **longest_only** - false

3.2.2 All Match and Best Match - Annotation with NOBLE Coder

NOBLE Coder was chosen against others similar tools because of it's comparable quality and higher ease of use. Each of the articles was annotated twice with this tool, using two different matching strategies, Best match and All match.

The commands used to annotate the reports were these:

```
$ java -jar NobleCoder-1.0.jar -terminology radlex \
-input [portuguese reports path] -output [output path] \
-search all-match\textit{
```

```
$ java -jar NobleCoder-1.0.jar -terminology radlex \
-input [portuguese reports path] -output [output path] \
-search best-match
```

[1] http://data.bioontology.org/documentation#nav_annotator

The RadLex ontology .owl file had to be edited before it could be correctly processed and uploaded to NOBLE Coder. In the original .owl file the properties "Preferred_name" and "Synonym" are considered to be *DatatypeProperty* but I had to change both to *AnnotationProperty*. That is, where in the file was

```
<owl:DatatypeProperty rdf:ID="Preferred_name">
</owl:DatatypeProperty>
```

I've had to change it to:

```
<owl:AnnotationProperty rdf:ID="Preferred_name">
</owl:AnnotationProperty>
```

And the analogous thing for the "Synonym" property.

3.3 Evaluation

The annotations of each MT or MT+PE translated article were compared against the annotations of corresponding HT translated article, which was considered a gold standard. Both Micro- and Macro- Precision, Recall and F1-scores were calculated. This was done for each matching approach.

To facilitate the understanding of the results, I will now walk trough a short example for one document. Consider that we have two Portuguese documents and for each one we have a HT English translation and a MT English translation. There were found 4 terms of interest in the HT translation, bone, cell, finger, colon. This is going to be our gold standard. In the MT translation, 2 terms of interest were found, brain, bone. One of these terms is also in the gold standard so there is 1 True Positive, but the other term is not, so that is 1 False Positive. In the gold standard there are 3 terms that were not found in the MT translation. That is 3 False Negatives. After calculations (see 2.4), this gives us a Precision score of 0.5, a Recall score of 0.25 and F-Score of 0.33.

These methods measure how similar are the terms annotated on the MT or MT+PE texts to the terms annotated on the HT texts. They don't say nothing about the quality of the annotations, however is that measured.

Chapter 4

Experimental Results

4.1 NER Gazetteer-based approach

Let's check the number of terms found by document.

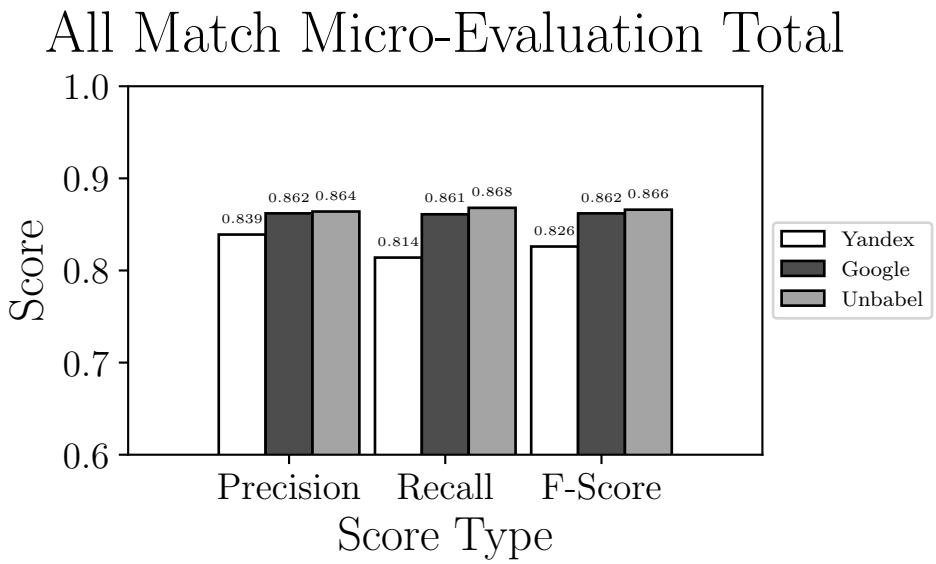
Translation	Direct Match	All Match	Best Match
Human	119.98	178.08	145.11
Yandex	115.34	172.66	144.19
Google	120.15	177.89	146.70
Unbabel	121.19	178.79	148.09

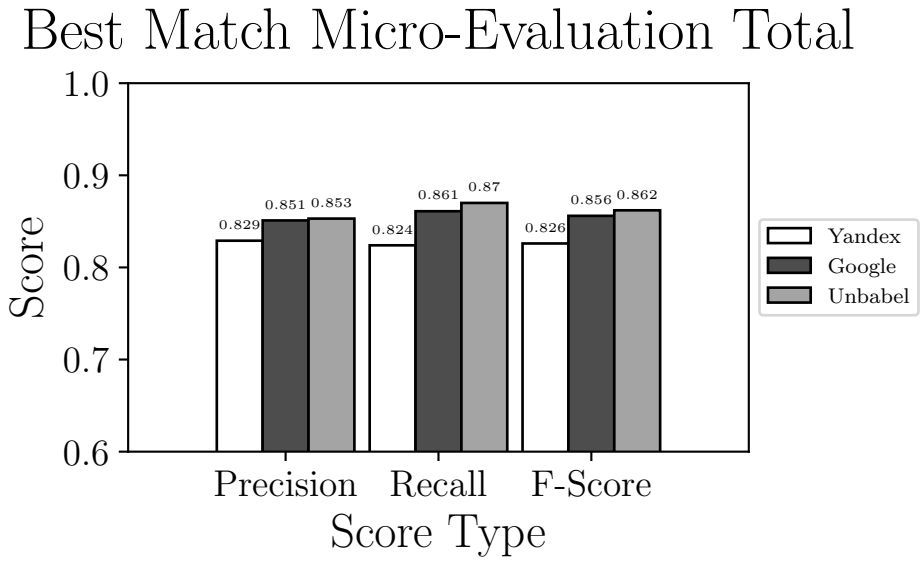
Table 4.1: Number of RadLex terms found by document

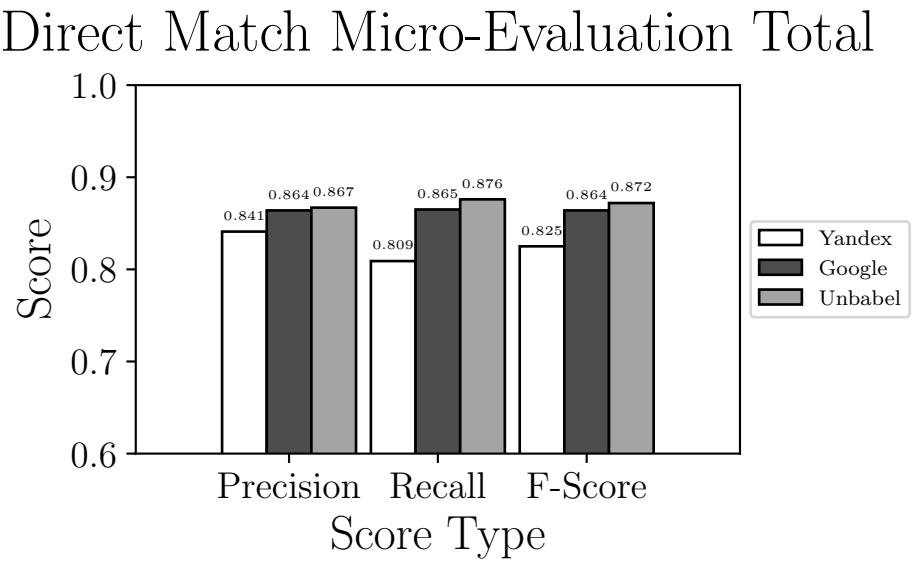
One of the highlights here is that the All Match approach consistently found more terms than the Best Match approach, which itself found more terms than the Direct Match approach. This makes sense since the All Match approach its the most liberal one in what it considers to be a mention of a RadLex term. The Best Match approach is more conservative than the All Match approach but less than the Direct Match approach, considering lexical variations and word reordering, for example. Let's now look at the differences between translations.

The highlights are that there were less RadLex terms in the Yandex translations than in any of the others. On average, there were less between 0.92 and 5.42 RadLex terms in each document, compared with the Human translations. On the other hand, on average there are more between 0.71 and 2.98 RadLex

terms in each Unbabel translation than in each Human translation. The Unbabel's translations were the ones with the most RadLex terms.







The annotations of the Google MT translation are closer to the ones of the HT translation than the ones of the Yandex MT translation. This could be just because the human translators used Google Translator to help them in their translation process, and so the translation is more similar than compared with the Yandex translation. This argument loses strength is we assume Google Translate translation outputs changed since the articles were human translated (publication years of the articles used range from 2003 to 2013), but I could not found data on this.

The annotations of Unbabel MT+PE translations are by itself closer to the ones of Human translations than the ones of Google translations. This is not surprising since the Post-Editing at Unbabel is done on the Google MT translation so we would expect the Unbabel results to the Human results than the Google results. The question is how much closer.

In the Introduction to thesis I've proposed the following hypothesis:

Hypothesis: MT+PE is a good trade-off between quality and cost, compared with MT and HT, for translating radiology reports for the purpose of identifying RadLex terms.

I've written that for this to be true, "MT+PE quality for the task at hand has to be close enough to HT quality". The MT+PE translation is indeed the

one closest to HT quality at the task at hand. This is not surprising, since that, from the types of translation being compared against the Human Translation, MT+PE is the only one which also involves humans in the translation process. But is it close enough to HT? This really would depend on the application of the annotations, but the results show that MT+PE translation results are far from being identical to the HT ones.

I also add the condition that for the hypothesis to be true, "MT+PE quality for the task at hand has to be better than MT quality, enough to compensate its higher cost". It is better indeed but if it is enough would also depend on the practical application of the annotations. Having said this, the results show that results from MT+PE translations are not very much closer to the results of HT translation comparing with Google MT translation. Again, how important this is would depend on the application of results of the NER task.

4.1.1 Clinical Finding Subset

Depending on the type of annotation approach and translation it was found between 15.75 and 22.58 "clinical finding" terms per document. For all three annotation strategies and both macro and micro evaluations, the terms extracted from the Yandex translation were the ones less similar to the ones extracted from the human translation.

In the micro evaluation, both Google and Unbabel were really similar, having equivalent F-scores in the Direct and Best Match Approach and Google having slightly higher F-Score score in the All-Match approach. In the macro evaluation, Unbabel had a slighter higher F-score in Direct and Best Match Approaches and equivalent to Google in the All Match approach.

4.1.2 Anatomical Entity Subset

Depending on the type of annotation approach and translation it was found between 19.15 and 32.38 "anatomical entity" terms per document. Scores similar to the total results in both Macro and Micro evaluations.

4.1.3 Imaging Modality Subset

Depending on the type of annotation approach and translation it was found between 2.64 and 3.62 "imaging modality" terms per document.

In the micro evaluation, Direct Match, approach the relative order in relation to F-Score are the same but the differences are really small. In the All Match approach, Yandex had the highest F-Score followed really closely by Unbabel and Google. In the Best Match approach, Yandex and Unbabel had the same scores and Google was the worst. For the macro evaluation, the results were similar although in in the Best Match Yandex was slightly (0.02 points) better than Unbabel)

4.1.4 Imaging Observation Subset

Depending on the type of annotation approach and translation it was found between 4.81 and 9.81 "imaging observation" terms per document. In the micro and macro evaluation, Google had better F-Scores than Unbabel and Yandex, except in the micro evaluation using a Best Match approach, in which it had the same F-Score as Unbabel.

4.1.5 Non-Anatomical Substance Subset

Depending on the type of annotation approach and translation it was found between 1.6 and 3.74 "Non-Anatomical Substance" terms per document.

In the micro evaluation, Unbabel had a better F-Score when Direct Match was used, Yandex when All Match was used and Google when Best Match was used. In the macro evaluation, when Direct Match was used the scores were lower than usual, around 0.65. Yandex had the best F-Score. Using other approaches, Google has a higher F-Score.

4.1.6 Object Subset

Depending on the type of annotation approach and translation it was found between 1.19 and 2.55 "Object" terms per document. In all cases Google had the highest F-Score.

4.1.7 Procedure Subset

Depending on the type of annotation approach and translation it was found between 5.36 and 7.38 "Procedure" terms per document.

4.1.8 Procedure Step

Depending on the type of annotation approach and translation it was found between 1.42 and 2.47 "Procedure Step" terms per document.

4.2 Discussion

4.3 Conclusions

Chapter 5

Conclusions

References

Aires, J., Lopes, G.P. & Gomes, L. (2016). English-Portuguese Biomedical Translation Task Using a Genuine Phrase-Based Statistical Machine Translation Approach. *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, **2**, 456–462. 15

Ananda-Rajah, M.R., Martinez, D., Slavin, M.A., Cavedon, L., Dooley, M., Cheng, A., Thursky, K.A., Ananda-Rajah, M., Cheng, A., Morrissey, C., Spelman, T., Dooley, M., Tomblyn, M., Chiller, T., Einsele, H., Gress, R., Sepkowitz, K., Yokoe, D., Casper, C., Dubberke, E., Lee, G., Munoz, P., Fourneret-Vivier, A., Lebeau, B., Mallaret, M., Brenier-Pinchart, M., Brion, J., Pauw, B.D., Walsh, T., Donnelly, J., Stevens, D., Edwards, J., Kontoyiannis, D., Marr, K., Park, B., Alexander, B., Anaissie, E., Lortholary, O., Gangneux, J., Sitbon, K., Lebeau, B., de Monbrison, F., Steinbach, W., Marr, K., Anaissie, E., Azie, N., Quan, S., Nicolle, M., Benet, T., Thiebaut, A., Bienvenu, A., Voirin, N., Neofytos, D., Treadway, S., Ostrander, D., Alonso, C., Dierberg, K., Ananda-Rajah, M., Grigg, A., Downey, M., Bajel, A., Spelman, T., Pagano, L., Caira, M., Candoni, A., Aversa, F., Castagnola, C., Denning, D., Mengoli, C., Cruciani, M., Barnes, R., Loeffler, J., Donnelly, J., Pfeiffer, C., Fine, J., Safdar, N., Maertens, J., Groll, A., Cordonnier, C., de la CÃ¡mara, R., Roilides, E., Chang, D., Burwell, L., Lyon, G., Pappas, P., Chiller, T., Marom, E., Kontoyiannis, D., Hota, B., Lin, M., Doherty, J., Borlawsky, T., Woeltje, K., Hazlehurst, B., Naleway, A., Mullooly, J., Elkin, P., Froehling, D.,

Wahner-Roedler, D., Brown, S., Bailey, K., Murff, H., FitzHenry, F., Matheny, M., Gentry, N., Kotter, K., Hripcsak, G., Friedman, C., Alderson, P., DuMouchel, W., Johnson, S., Elkins, J., Friedman, C., Boden-Albala, B., Sacco, R., Hripcsak, G., Cooley, L., Spelman, D., Thursky, K., Slavin, M., Morrissey, C., Chen, S., Sorrell, T., Milliken, S., Bardy, P., Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I., Cohen, K., Hunter, L., Stone, M., Goldman, R., Pagano, L., Caira, M., Candoni, A., Offidani, M., Martino, B., Altman, D., Bland, J., Hripcsak, G., Kuperman, G., Friedman, C., Heitjan, D., Fiszman, M., Chapman, W., Aronsky, D., Evans, R., Haug, P., Haas, J., Mendonca, E., Ross, B., Friedman, C., Larson, E., Azie, N., Neofytos, D., Pfaller, M., Meier-Kriesche, H., Quan, S., D'Avolio, L., Nguyen, T., Farwell, W., Chen, Y., Fitzmeyer, F., Wang, Z., Shah, A., Tate, A., Denaxas, S., Shawe-Taylor, J., Ananda-Rajah, M., Slavin, M. & Thursky, K. (2014). Facilitating Surveillance of Pulmonary Invasive Mold Diseases in Patients with Haematological Malignancies by Screening Computed Tomography Reports Using Natural Language Processing. *PLoS ONE*, **9**, e107797. 7

Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *To appear: EMNLP-2016*. 8

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Herve, S.A., Soricut, R., Specia, L. & Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. *2014 Workshop on Statistical Machine Translation*, 12–58. 14

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K. & Zampieri, M.

(2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, vol. 2, 131–198. 15

Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B.A., Santaholma, M.E., Starlander, M., Nakao, Y., Kanzaki, K. & Isahara, H. (2005). A Generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, 50–58. 13

Bretschneider, C., Oberkampf, H., Zillner, S., Bauer, B. & Hammon, M. (2014). Corpus-based Translation of Ontologies for Improved Multilingual Semantic Annotation. *Coling*, 1–8. 2, 16

Castilla, C. (2007). Instrumento de Investigação Clínico-Epidemiológica em Cardiologia Fundamentado no Processamento de Linguagem Natural. 2, 12, 15

A survey of current work in biomedical text mining. 5

Cotik, V., Filippo, D. & Castaño, J. (2015). An Approach for Automatic Classification of Radiology Reports in Spanish. *Studies in Health Technology and Informatics*, **216**, 634–638. 7

Dang, P.A., Kalra, M.K., Schultz, T.J., Graham, S.A. & Dreyer, K.J. (2009). Informatics in radiology: Render: an online searchable radiology study repository. *Radiographics : a review publication of the Radiological Society of North America, Inc*, **29**, 1233–46. 7

Do, B.H., Wu, A., Biswal, S., Kamaya, A. & Rubin, D.L. (2010). Informatics in Radiology: RADTF: A Semantic Search–enabled, Natural Language Processor–generated Radiology Teaching File. *RadioGraphics*, **30**, 2039–2048. 17

Dreyer, K.J., Kalra, M.K., Maher, M.M., Hurier, A.M., Asfaw, B.A., Schultz, T., Halpern, E.F. & Thrall, J.H. (2005). Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports: Validation Study. *Radiology*, **234**, 323–329. 6, 7

REFERENCES

Durrani, N., Haddow, B., Koehn, P. & Heafield, K. (2014). Edinburgh's Phrase-based Machine Translation Systems for WMT-14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 97–104. 14

Eck, M., Vogel, S. & Waibel, A. (2004). Improving statistical machine translation in the medical domain using the unified medical language system. *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, 792–es. 14

Elhadad, N., Pradhan, S., Gorman, S.L., Manandhar, S., Chapman, W.W. & Savova, G. (2015). SemEval-2015 Task 14 : Analysis of Clinical Text. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 303–310. 20

G, R., M, F., R, A., O, E. & K, P. (2013). Using machine translation in clinical practice. *Canadian family physician Medecin de famille canadien*, **59**, 382–383. 14

Gerstmair, A., Daumke, P., Simon, K., Langer, M. & Kotter, E. (2012). Intelligent image retrieval based on radiology reports. *European Radiology*, **22**, 2750–2758. 2, 17

Hassanpour, S. & Langlotz, C.P. (2016). Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, **66**, 29–39. 1, 7

Hong, Y. & Kahn, C.E. (2013). Content analysis of reporting templates and free-text radiology reports. *Journal of Digital Imaging*, **26**, 843–849. 7

Hong, Y., Zhang, J., Heilbrun, M.E. & Kahn, C.E. (2012). Analysis of RadLex coverage and term co-occurrence in radiology reporting templates. *Journal of Digital Imaging*, **25**, 56–62. 17

Hotho, A., Nürnberger, A. & Paass, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, **20**, 19–62. 5

Jonquet, C., Shah, N.H. & Musen, M.A. (2009). The open biomedical annotator. *Summit on translational bioinformatics*, **2009**, 56–60. 6, 18

Kaliyadan, F. & Pillai, S.G. (2010). The use of Google language tools as an interpretation aid in cross-cultural doctor-patient interaction: A pilot study. *Informatics in Primary Care*, **18**, 141–143. 13, 14

Kathol, A., Precoda, K., Vergyri, D., Wang, W., Riehemann, S., International, S.R.I. & Park, M. (2005). Speech Translation for Low-Resource Languages : The Case of Pashto. *Syntax*, 2273–2276. 14

Kirchhoff, K., Turner, A.M., Axelrod, A. & Saavedra, F. (2011). Application of statistical machine translation to public health information: a feasibility study. *Journal of the American Medical Informatics Association : JAMIA*, **18**, 473–478. 14

Klein, J.S. (2013). A Look Back at 2012 and Plans for 2013. *RadioGraphics*, **33**, 1–2. 17

Koehn & Philipp (2010). Statistical Machine Translation. *Cambridge: Cambridge University Press*, 433. 8

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *Journal of Specialised Translation*, 131–148. 11

Kurtz, C., Depeursinge, A., Napel, S., Beaulieu, C.F. & Rubin, D.L. (2014). On combining image-based and ontological semantic dissimilarities for medical image retrieval applications. *Medical Image Analysis*, **18**, 1082–1100. 17

Langlotz, C.P. (2006). RadLex: a new method for indexing online educational materials. *Radiographics : a review publication of the Radiological Society of North America, Inc*, **26**, 1595–1597. 16

REFERENCES

Langlotz, C.P. & Caldwell, S.A. (2002). The completeness of existing lexicons for representing radiology report information. *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology*, **15 Suppl 1**, 201–205. 16

Mansouri, A., Affendey, L. & Mamayt, A. (2008). Named Entity Recognition Approaches :. *IJCSNS International Journal of Computer Science and Network Security*, **8**. 6

Marta R. Costa-jussà, Mireia Farrús, J.S.P. (2012). Machine Translation in Medicine. In *ARSA - PROCEEDINGS IN ARSA - ADVANCED RESEARCH IN SCIENTIFIC AREAS*, 1995–1998, EDIS - Publishing Institution of the University of Zilina. 14

Martinez, D., Ananda-Rajah, M.R., Suominen, H., Slavin, M.A., Thursky, K.A. & Cavedon, L. (2015). Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. *Journal of Biomedical Informatics*, **53**, 251–260. 7

Marwede, D., Schulz, T. & Kahn, T. (2008). Indexing thoracic CT reports using a preliminary version of a standardized radiological lexicon (RadLex). *Journal of Digital Imaging*, **21**, 363–370. 16

Musleh, A., Durrani, N., Temnikova, I., Nakov, P., Vogel, S. & Alsaad, O. (2016). Enabling Medical Translation for Low-Resource Languages. *Proceedings of the 16th Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*. 14

Nagata, M. & Pedersen, T. (2005). Proceedings of the ACL Interactive Poster and Demonstration Sessions. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. 13

Patil, S. & Davies, P. (2014). Use of Google Translate in medical communication: evaluation of accuracy. *BMJ (Clinical research ed.)*, **349**, g7392. 13

Pons, E., Braun, L.M.M., Hunink, M.G.M. & Kors, J.A. (2016). Natural Language Processing in Radiology: A Systematic Review. *Radiology*, **279**, 329–343. 1, 5

Robinson, P.N. & Bauer, S. (2011). *Introduction to bio-ontologies*. CRC Press. 7

Schyve, P.M. (2007). Language differences as a barrier to quality and safety in health care: The joint commission perspective. *Journal of General Internal Medicine*, **22**, 360–361. 13

Somers, H. (2003). *Computers and translation: a translator's guide*. John Benjamins Publishing Company. 11

Spanier, A.B., Cohen, D. & Joskowicz, L. (2016). A new method for the automatic retrieval of medical cases based on the RadLex ontology. *International Journal of Computer Assisted Radiology and Surgery*. 17

Stewart, S.A., Von Maltzahn, M.E. & Abidi, S.S.R. (2012). Comparing metamap to mgrep as a tool for mapping free text to formal medical lexicons. In *CEUR Workshop Proceedings*, vol. 895, 63–77. 6, 20

Tatsumi, M. & Aikawa, T. (2012). How Good Is Crowd Post-Editing? Its Potential and Limitations. *... 2012 Workshop on ...*. 11

Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G. & Jacobson, R.S. (2016). NOBLE - Flexible concept recognition for large-scale biomedical natural language processing. *BMC bioinformatics*, **17**, 32. xxi, 6, 18, 19, 20

Turner, A.M., Dew, K.N., Desai, L., Martin, N. & Kirchhoff, K. (2015). Machine Translation of Public Health Materials From English to Chinese: A Feasibility Study. *JMIR public health and surveillance*, **1**, e17. 14

Wang, K.b., Salunkhe, A., Morriso, J., Lee, P., Mejino, J., Detwiler, L.f., Brinkley, J.F., Siegel, E.d., Rubin, D. & Carrino, J. (2015). Ontology-based image navigation: Exploring 3.0-T MR neurography of the brachial plexus using AIM and radlex. *Radiographics*, **35**, 142–151. 17

REFERENCES

Wołk, K. & Marasek, K. (2015). Neural-based Machine Translation for Medical Text Domain. Based on European Medicines Agency Leaflet Texts. In *Procedia Computer Science*, vol. 64, 2–9. 14

Woods, R.W. & Eng, J. (2013). Evaluating the completeness of radlex in the chest radiography domain. *Academic Radiology*, **20**, 1329–1333. 16

Zeng-Treitler, Q., Kim, H., Rosemblat, G. & Keselman, A. (2010). Can multilingual machine translation help make medical record content more comprehensible to patients? In *Studies in Health Technology and Informatics*, vol. 160, 73–77. 15