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

Dissertação orientada por:
Prof. Doutor Francisco Couto
Dr. Vasco Pedro

2017

Resumo

Os relatórios de Radiologia descrevem os resultados dos procedimentos de radiografia e têm o potencial de ser uma fonte de informação útil que pode trazer benefícios para os sistemas de saúde ao redor do mundo. No entanto, estes relatórios são geralmente escritos em texto livre e, portanto, é difícil extrair automaticamente informação a partir deles. Contudo, o fato de que a maioria dos relatórios estão agora digitalmente disponíveis torna-os passíveis de utilização de ferramentas de Prospeção de Texto (*Text Mining*). Outra vantagem dos relatórios de Radiologia, que os torna mais suscetíveis à utilização destas ferramentas, é que mesmo se escritos em texto livre, eles são geralmente bem estruturados. O problema é que estas ferramentas são principalmente desenvolvidas para Inglês e os relatórios são geralmente escritos na língua nativa do radiologista, que não é necessariamente o Inglês. Isso cria um obstáculo para a partilha de informação de Radiologia entre diferentes comunidades, partilha esta importante para compreender e tratar eficazmente problemas de saúde.

Existem basicamente duas soluções possíveis para este problema. Se se aplicar, uma solução é traduzir o próprio léxico que é utilizado pela ferramenta de Prospeção de Texto que se pretende utilizar. A outra é traduzir os próprios relatórios. Traduzir o léxico tem a vantagem de não necessitar de tradução contínua, ou seja, depois de traduzir um léxico para, por exemplo, Espanhol, podemos usá-lo para processar tantos relatórios Espanhóis não traduzidas conforme necessário. No entanto, quando uma nova versão do léxico é lançada as mudanças também precisam de ser traduzidas, caso contrário, o léxico traduzido ficaria desatualizado. Dada a crescente evolução de serviços de tradução hoje disponíveis, neste trabalho é avaliada a opção alternativa de traduzir os relatórios e verificar a sua viabilidade. Esta

abordagem tem a vantagem de que os relatórios traduzidos seriam acessíveis a qualquer médico que entenda Inglês e quaisquer ferramentas estado da arte de Prospeção de Texto focadas em texto em Inglês podem ser aplicadas sem qualquer necessidade de adaptação.

Se a tradução for feita por profissionais treinados em tradução de textos médicos, provavelmente pode-se assumir que informação não se perde no processo de tradução. Chamamos a este tipo de tradução Tradução Humana (*Human Translation*). Mas a utilização de tradutores especializados é cara, o que faz com que esta solução não seja escalável. Outra opção é usar Tradução Automática (*Machine Translation*). Não obstante a menor qualidade da tradução, é mais barata e mais viável em grande escala. Finalmente, uma opção que tenta obter o melhor dos dois mundos é usar Tradução Automática seguida de Pós-Edição (*Post-Edition*) por humanos. Nesta abordagem, o texto é automaticamente traduzido e, em seguida, a tradução é corrigida por um humano. Mais barata do que a opção de Tradução Humana e com melhor qualidade do que a de Tradução Automática.

A escolha de abordagem de tradução é importante porque vai afetar a qualidade dos resultados das ferramentas de Prospeção de Texto. Atualmente não há nenhum estudo disponível publicamente que tenha fornecido evidência quantitativa que auxilie a fazer esta escolha. Isto pode ser explicado pela falta de um corpus paralelo que poderia ser usado para estudar este problema.

Este trabalho explora a solução de traduzir os relatórios para Inglês antes de aplicar as ferramentas de Prospeção de Texto, analisando a questão de qual a abordagem de tradução que deve ser usada. Com este fim, criei MRRAD (*Multilingual Radiology Research Articles Dataset*), um corpus paralelo de 51 artigos portugueses de investigação relacionados com Radiologia, e uma série de traduções alternativas (humanas, automáticas e semi-automáticas) para Inglês. As versões originais dos artigos, em Português, e as traduções humanas foram extraídas automaticamente da biblioteca online SciELO.

Traduções automáticas foram obtidas utilizando os serviços da Yandex e da Google e traduções semi-automáticas através dos serviços da Unbabel. Este é um corpus original que pode ser usado no avanço da investigação sobre este tema.

Usando o MRRAD estudei que tipo de abordagem de tradução automática ou semi-automática é mais eficaz na tarefa de Reconhecimento de Entidades (*Named-Entity Recognition*) RadLex na versão em Inglês dos artigos. RadLex é uma ontologia que se foca em termos relacionados com Radiologia. A tarefa de Reconhecimento de Entidades é relevante uma vez que os seus resultados podem ser usadas em sistemas de Recuperação de Imagens (*Image Retrieval*) e de Recuperação de Informação (*Information Retrieval*) e podem ser úteis para melhorar Sistemas de Respostas a Perguntas *(Question Answering)*. Para realizar o Reconhecimento de Entidades RadLex utilizei a API do Open Biomedical Annotator e duas diferente configurações do software NOBLE Coder. Assim, ao todo utilizei três diferentes abordagens para identificar termos RadLex nos textos. A diferença entre as abordagens está em quão flexíveis ou estritas estas são em identificar os termos.

Considerando os termos identificados nas traduções humanas como o padrão ouro (*gold-standard*), calculei o quão semelhante a este padrão foram os termos identificados usando outras abordagens de tradução. Descobri que uma abordagem completamente automática de tradução utilizando o Google leva a micro F-Scores (entre 0,861 e 0,868, dependendo da abordagem de reconhecimento) semelhantes aos obtidos através de uma abordagem mais cara, tradução semi-automática usando Unbabel (entre 0,862 e 0,870). A abordagem de tradução utilizando os serviços da Yandex obteve micro F-Scores mais baixos (entre 0,829 e 0,831). Os resultados foram semelhantes mesmo no caso onde se consideraram apenas termos de RadLex pertences às sub-árvores correspondentes a entidades anatómicas e achados clínicos.

Para entender melhor os resultados, também realizei uma análise qualitativa do tipo de erros encontrados nas traduções automáticas e semi-automáticas. A análise foi feita sobre os Falsos Positivos (FPs) e Falsos Negativos (FNs) cometidos pelas traduções utilizando Yandex, Google e Unbabel em 9 documentos aleatórios e cada erro foi classificado por tipo. A maioria dos FPs e FNs são explicados não por uma tradução errada mas por outras causas, por exemplo, uma tradução alternativa que leva a uma diferença nos termos identificados.

Poderia ser esperado que as traduções Unbabel tivessem muitos menos erros, visto que têm o envolvimento de humanos, do que as da Google, mas isso nem sempre acontece. Há situações em que erros são até adicionados mesmo durante a etapa de Pós-Edição. Uma revisão dos erros faz-me propor que isso poderá ser devido à falta de conhecimento médico dos editores (utilizadores responsáveis por fazer a Pós-Edição) atuais da Unbabel. Por exemplo, um *stroke* (acidente vascular cerebral) é algo que ocorre no cérebro, mas num caso foi usado como algo que acontece no coração - alguém com algum conhecimento sobre a medicina não faria este erro. Mas a verdade é que a Unbabel atualmente não se foca em conteúdo médico. Prevejo que se eles o fizessem e investissem em crescer uma comunidade de utilizadores especialistas com melhor conhecimento da linguagem médica, isso levaria a melhores resultados.

Dito isto, os resultados deste trabalho corroboram a conclusão de que se desenvolvedores de *software* tiverem recursos financeiros limitados para pagar por Tradução Humana, ficarão melhor servidos se usarem um serviço de tradução automática como a Google em vez de um serviço que implementa Pós-Edição, como a Unbabel. É claro que talvez haja melhores serviços de Tradução Automática do que a Google ou melhores serviços de Tradução Automática + Pós-Edição do que a Unbabel oferece atualmente para o campo médico, e isso é algo que poderia ser explorado em trabalhos futuros.

O corpus MRRAD e as anotações utilizadas neste trabalho podem ser encontradas em https://github.com/lasigeBioTM/MRRAD.

Palavras Chave: Tradução, Reconhecimento de Entidades, Corpus Paralelo, Radiology, RadLex

Abstract

Radiology reports describe the results of radiography procedures and have the potential of being an useful source of information which can bring benefits to health care systems around the world. One way to automatically extract information from the reports is by using Text Mining tools. The problem is that these tools are mostly developed for English and reports are usually written in the native language of the radiologist, which is not necessarily English. This creates an obstacle to the sharing of Radiology information between different communities.

This work explores the solution of translating the reports to English before applying the Text Mining tools, probing the question of what translation approach should be used. Having this goal, I created MR-RAD (Multilingual Radiology Research Articles Dataset), a parallel corpus of Portuguese research articles related to Radiology and a number of alternative translations (human, automatic and semi-automatic) to English. This is a novel corpus which can be used to move forward the research on this topic.

Using MRRAD I studied which kind of automatic or semi-automatic translation approach is more effective on the Named-entity recognition task of finding RadLex terms in the English version of the articles. Considering the terms identified in human translations as the gold standard, I calculated how similar to this standard were the terms identified using other translation approaches (Yandex, Google and Unbabel). I found that a completely automatic translation approach using Google leads to micro F-Scores (between 0.861 and 0.868, depending on the identification approach) similar to the ones obtained through a more expensive semi-automatic translation approach using Unbabel (between 0.862 and 0.870). To better understand the

results I also performed a qualitative analysis of the type of errors found in the automatic and semi-automatic translations. The MR-RAD corpus and annotations used in this work can be found at https://github.com/lasigeBioTM/MRRAD.

Keywords: Translation, Named-entity Recognition, Parallel Corpus, Radiology, RadLex

Acknowledgements

I want to thank my supervisors for the availably and opportunity. I also want to show my appreciation for the members of the Unbabel's team who gave me technical support with their API.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Radiology reports describe the results of radiography procedures and have the potential of being an useful source of information, which can bring benefits to health care systems around the world. However, these reports are usually written in free-text and thus it is hard to automatically extract information from them. Nonetheless, the fact that most reports are now digitally available make them amenable for using Text Mining tools. Another advantage of Radiology reports is that even if written in free-text, they are usually well structured.

A lot of work has been done on Text Mining of Biomedical texts, including health records (Pons *et al.*, 2016), but although Radiology reports are usually written in the native language of the radiologist, Text Mining tools are mostly developed for 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. Given this dependence, the system cannot be easily applied to reports written in other languages. And even if the system was not dependable on an English lexicon, it is not certain that the results would be the same if another language was used, because of, for example, differences in syntax. This have been an obstacle in the sharing of Radiology information between different communities, which is important to understand and effectively address health problems.

There are mainly two possible solutions to this problem. One is to translate the lexicon itself (Bretschneider *et al.*, 2014) and the other is to translate the reports. Translating the lexicon has the advantage of not requiring continuous translation, i.e., after translating a lexicon to, for example, Spanish, we can then use it to process as many untranslated Spanish reports as needed. However, when a new version of the lexicon is released the changes need also to be translated, otherwise the translated lexicon would become outdated. Given the increasing evolution of translation services nowadays available, in this work I assess the alternative option of translating the reports and check its feasibility. This approach has the advantage that the translated reports would be accessible to any doctor who understands English and any state-of-the-art Text Mining tools focused on English text can be applied without any need for adaptation.

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 expert translators are expensive, which makes this solution unscalable, with a high financial 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. In this approach the text is automatically translated and then the translation is corrected by a human. Cheaper than the HT option and with better quality than the MT one.

The choice of translation approach its important because it will affect the quality of the output of the Text Mining tools. To the best of my knowledge, currently there is no publicly available study that provided a quantitative evidence that would help make this choice. This could be explained by the lack of a parallel corpus that could be used to study this. 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 2.3.5.1). The author does not compare translation systems, something that is done on the present work.

Specifically, I focused on the Text Mining task of Named-entity recognition (NER). This is a relevant task since the outputs from NER systems can be used in Image Retrieval (Gerstmair *et al.*, 2012) and Information Retrieval (Antony & Suryanarayanan Mahalakshmi, 2015) systems and can be useful for improving automatic Question Answering (Toral *et al.*, 2005).

1.2 Objectives

Thus, I aimed at addressing the following research question: lacking the resources to pay for Human Translation services, what kind of automatic (MT) or semi-automatic translation (MT+PE) approach should be used in the task of translating Portuguese Radiology-related text to English, for the purposes of finding RadLex terms in the translated text? I propose the following hypothesis:

Hypothesis: MT+PE is a good trade-off between quality and cost, compared with MT and HT, for translating Portuguese Radiology reports to English, for the purpose of identifying RadLex terms in the translated text.

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

1. MT+PE has to be cheaper than HT
2. The terms identified in the MT+PE translations have to be similar enough to the ones identified in the HT translation
3. The terms identified in MT+PE translations have to be more similar to the ones identified in the HT translation than the ones identified in MT translations

The first condition is known to be true. The last condition its important because if MT+PE quality is similar to MT quality, as MT cost is lower, then it is 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 test this hypothesis I have compared the RadLex terms identified in MT and MT+PE translations to the terms identified in HT translations, which I assumed to be a gold standard.

For this purposes I've created MRRAD, a parallel corpus containing 51 Portuguese scientific articles related to Radiology and corresponding HT, MT (Google and Yandex) and MT+PE (Unbabel) English translations. These translations were annotated with RadLex terms using the Open Biomedical Annotator and NOBLE Coder. More than one annotation approach was used to experiment with different kinds annotation approaches. For each translation and annotation approach I created the set of the RadLex terms that were found in that translations with that annotation approach. The terms found in the MT and MT+PE translations were then compared with the ones found in the HT translations.

The MRRAD corpus and annotations used in this work can be found in a public GitHub repository[1].

1.4 Contributions

This thesis lead to the following contributions:

- **MRRAD Corpus**
  - A Portuguese-English parallel corpus of research articles related to Radiology, called MRRAD (Multilingual Radiology Research Articles Dataset), containing for each article the original Portuguese document, the HT translation, two alternative MT translations and a MT+PE translation;
- **Main Scientific Results**

[1] `https://github.com/lasigeBioTM/MRRAD`

  - Measurement of the performance of multiple automatic or semi-automatic translation approaches in the task of translating Portuguese Radiology-related text to English, for the purposes of finding RadLex terms in the translated text;

- **Bioinformatics Open Days 2017**[1]
  - Abstract submission and oral presentation describing this work;
  - Organization and presentation of workshop on Biomedical Text Mining with other members of the LaSIGE team[2];
- **Scientific Publications**
  - ...
- **Other Open-Source Contribuitions**
  - A Python binding of the BioPortal REST API[3];
  - Converter of NOBLE Coder annotation file to Webanno TSV 2 annotations files[4];
- **Multilingual Report Annotator**
  - Development of a proof of concept web application for translation and annotation of Radiology text;[5]

[1] http://bioinformaticsopendays.com/
[2] https://sites.google.com/view/biomedicaltextminingworkshop
[3] https://github.com/LLCampos/pybioportal
[4] https://gist.github.com/LLCampos/5f1680941984c4b63f986965e7384e6c
[5] http://www.lasige.di.fc.ul.pt/webtools/mra/

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).

2.1.1 Named-entity Recognition

Named-entity recognition (NER) is a task of Text Mining 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, there are NER systems that can identify mentions of chemical entities (Zhang *et al.*, 2016), diseases (Wei *et al.*, 2016) or terms from specific ontologies like HPO (Human Phenotype Ontology) (Groza *et al.*, 2015). Considering the case of diseases, in the phrase *Atrial fibrillation has strong associations with other cardiovascular diseases* the term *Atrial fibrillation* is a named-entity that represents a disease. This is a relevant task since the outputs from NER systems can be used in Image Retrieval (Gerstmair *et al.*, 2012) and Information Retrieval (Antony & Suryanarayanan Mahalakshmi, 2015) systems and can be useful for improving automatic Question Answering (Toral *et al.*, 2005).

The approaches of NER can be divided into three categories (Mohit, 2014): 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 Machine Learning 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.

Lexicon based-approaches are a subset of the rule-based approaches. In this approach we already have a list of the named-entities (a lexicon) that we want to identify in the text. For example, if we want to identify chemical entities in text, we use a chemical-entities lexicon. The goal of the lexicon based-approach is then to identify, in text, mentions of terms presented in the lexicon. 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 lexicon, not considering, for example, lexical variations. The recall can be lower than expected because lexical variants (like plurals), abbreviations and partial matchings of lexicon terms are not identified 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 Natural Language Processing

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 sections I will briefly explain some of the NLP tasks relevant to this thesis.

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

2.1.2.1 Tokenization

Tokenization is one of the main tasks of NLP and consists in dividing a certain text in pieces called *tokens*. A token can be defined as "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing" (Manning *et al.*, 2009c). So, for the sentence *the mother had a surgery*, it is possible to divide it in five tokens, one for which word, using the heuristic that each token is separated by a whitespace. For more complicated text, one could intuitively think that a good strategy would be to split on all non-alphanumeric characters, but this sometimes raises problems. This strategy would tokenize *isn't* in *isn* and *t* which is intuitively wrong. More complicated strategies are needed. Relevant to this work, the tokenization strategies used are language specific. For example, one should not use an English tokenizer to tokenize Portuguese text (Branco & Silva, 2003).

2.1.2.2 Stemming and Lemmatization

Sometimes it is necessary to normalize lexical variations of a word to a base form, e.g., normalize the words *car*, *cars*, *car's* and *cars'* to just *car*. This can be useful, for example, in lexicon-based NER applications. If the the word *car* in the lexicon, it can make sense to consider lexical variations of the word *car* to be mentions of *car*. This can be accomplished by normalizing the words in the lexicon and in the text. This is done by using one of two techniques, Stemming or Lemmatization (Manning *et al.*, 2009b). In Stemming, crude rules are applied to cut off the suffixes of a word, the most popular stemmer being Porter's algorithm (Porter, 1980). On the other hand, Lemmatization does something similar but considers the context of the word.

2.1.3 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 on information theory to classify reports as having or not having important clinical findings and as having or 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 its 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 the relevance of the retrieved information.

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

2.1.4 Ontologies

To answer the questions presented in Chapter 1, the RadLex ontology is used. 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 a 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 Datasets and Corpora

Since one of the main goals of this thesis was to study Radiology reports, I did research on the available relevant datasets/corpora. Although I found a lot of public accessible Radiology documents, translations were not available and so they were not used in the work leading to this dissertation. A briefly description of each of the datasets/corpora found is presented next.

2.2.1 MIMIC II Clinical Database and MIMIC III Critical Care Database

The MIMIC II Clinical Database[4] is one of the MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) Databases. This dataset contains clinical data on tens of thousands of patients in Intensive Care Units, collected between 2001 and 2008. The data includes a number of procedures reports, including Radiology reports.

In August of 2015, a extension of MIMIC II was launched, called MIMIC III[5] (Medical Information Mart for Intensive Care III)(Johnson *et al.*, 2016), containing new data collected between 2008 and 2012.

2.2.2 Lurie Children's Teaching File Library

The Medical Imaging Resource Community (MIRC) is an open-source project which aims to develop free software tools for education and research in Radiology. Lurie Children's Hospital of Chicago[6] makes use of one of these tools, the Teaching

[1] http://www.geneontology.org/
[2] http://www.snomed.org/snomed-ct
[3] https://www.ebi.ac.uk/chebi/
[4] https://physionet.org/mimic2/mimic2_clinical_overview.shtml
[5] https://mimic.physionet.org/
[6] https://www.luriechildrens.org/en-us/Pages/index.aspx

Files System (TFS), to make available[1], for education purposes, more than 2,000 Radiology reports accompanied with corresponding Radiology images.

2.2.3 iDASH - Clinical Notes and Reports

iDASH[2] openly provides 2,363 medical transcription samples, including Radiology reports, extracted from Medical Transcription Samples website[3].

2.3 Translation

2.3.1 Terminology

During this dissertation some terminology related to translation practices is used. In this section I briefly explain this terms (Koehn & Philipp, 2010).

Parallel Corpora - A *corpus* is just a set of texts (*corpora* is used if you want to refer to more than one of these sets). The term *parallel corpus* is used to refer to a set of texts paired with corresponding translations into other languages.

Language Pair - This term refers to the languages involved in a translation. For example, in a translation from Portuguese to English, we can say that the language pair is Portuguese-English, Portuguese being the *source language* and English the *target language*.

2.3.2 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 its 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).

[1] http://mirc.luriechildrens.org/query
[2] https://idash.ucsd.edu/
[3] urlhttp://www.medicaltranscriptionsamples.com

2.3.2.1 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 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 does not make sense but the computer does not 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 for every word in the source text.

Table 2.1: Lexical translation probability table for the word broken

broken

t	p(t|s)
partido	0.8
falido	0.2

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 what is called 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 do not 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 (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 does not 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). $\Pr(s|t)$ 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.

[1] *red* can be translated to *vermelho* and *swelling* to *inchaço*

2.3.2.2 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.

2.3.3 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 cannot 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 this 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 *et al.*, 2012). The advantages of this strategy include lower per-word cost and sometimes an higher speed, compared with HT. One big disadvantage is less assurance of quality.

2.3.4 Machine Translation Services

2.3.4.1 Yandex

Yandex[1] is a Russian search-engine company. Currently, 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 could not find any research paper evaluating the translation's quality of Yandex.Translate in the language pair Portuguese-English.

2.3.4.2 Google

Google[3] is a company from the United States that sells a lot of technological services, including Machine Translation. For the language pair Portuguese-English, their translation services now uses Neural Machine Translation[4] (see section 2.3.1), although it is still possible to obtain Statistical based translations through their API.

2.3.4.3 Unbabel

Unbabel[5] is a Portuguese start-up which sells translation services focused on conversational content like costumer service or website copywriting, using an MT+PE approach. Although it is not mentioned in the Unbabel's API documentation, for the language pair Portuguese-English, Unbabel currently uses Google Translate's services in the 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);

[1] `https://yandex.com/`

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

[3] `https://www.google.com`

[4] `https://blog.google/products/translate/found%2Dtranslation%2Dmore%2Daccurate%2Dfluent%2Dsentences%2Dgoogle%2Dtranslate`

[5] `https://unbabel.com/`

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 am going to call this type of translation *Unbabel Translation*.

2.3.5 Translation of Medical Text

2.3.5.1 Multilingual Text Mining

There is not much research studying the effect of translation on Text Mining tools. (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 with a specialized medical translation dictionary. 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.

2.3.5.2 Machine Translation of Doctor-Patient Communication

Most of the work I found 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 do not 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 any other system. 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 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 (Kaliyadan & Pillai, 2010; Marta R. Costa-jussà, Mireia Farrús, 2012; Randhawa *et al.*, 2013) 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).

2.3.5.3 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.

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

[2] `http://www.universaldoctor.com/`

2.3.5.4 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].

2.3.5.5 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 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[3] 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

[1] `http://khresmoi.eu/`
[2] `http://www.statmt.org/moses/`
[3] `https://www.babelfish.com/`

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 submissions 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.

2.3.6 Translation of Biomedical Lexicons

One alternative solution to the one I am 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 has a Spanish corpus and wants to annotate it with terms of some lexicon, 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. This example is similar to (Cotik *et al.*, 2015), in which all RadLex terms were translated to Spanish using Google Translate and medical reports were annotated with this translated terms.

For the German language there is (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. They do this semi-automatically with the help of the corpus they wanted to annotate. With this, the authors improved the annotation of German text with RadLex terms.

2.4 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.

2.4.1 RadLex

RadLex[1] is an 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 8.000 to around 75.000 in just ten years. Being an ontology, RadLex can be visualized as a tree, which contains other subtrees. This characteristic can be used to extract subsets of the RadLex ontology. For example, if someone just wants to use the RadLex classes related to clinical findings she could just use the RadLex subtree containing just the children of the RadLex class *clinical finding*.

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 inclusion in the study of categories such as *procedures*, which did not had any match with RadLex terms. They also used a different methodology to find matches between manually 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

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

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 are not necessarily relevant anymore.

One could use RadLex to assist in the matching of research 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).

2.4.2 Open Biomedical Annotator

The Open Biomedical Annotator (OBA)[1] is an open-source tool for NER using a lexicon-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. OBA uses MGrep, which implements a radix-tree based data structure that allows for a fast

[1] `http://bioportal.bioontology.org/annotator`

match between terms in a lexicon and terms in text. OBA 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).

2.4.3 NOBLE Coder

NOBLE Coder[1] (Tseytlin *et al.*, 2016) is a software for NER using a lexicon-based approach. The lexicon is set by the user (it has to be in UMLS (RRF)[2], OWL[3] or OBO[4] formats or be present in BioPortal[5]). The lexicon is processed into two hash-tables which are then used during by NOBLE to find, in an arbitrary text, mentions of terms found in the lexicon.

Unlike the system used by OBA, NOBLE can find mentions of lexical variations of the terms present in the lexicon because it applies word Stemming. For example, *lobe* is a term present in the RadLex terminology, but its plural, *lobes*, is not. 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 lexicon 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 match the longest mention. For example, *toe*, *toe skin* and *skin* are all RadLex terms. If the "Subsumption" option is set, in the text *toe skin*, only the term *toe skin* will be identified. Otherwise, the terms *toe* and *skin* are also identified.

- **Overlap** - If this option is used, matched terms can overlap each other. For example, if this option is not set, NOBLE will only identify the terms *deep* and *lateral margin* in the text *deep lateral margin*. If it is set, it will also regnize the term *deep margin* which overlaps with the two other terms.

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

- **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. Its possible to set how many irrelevant words can be between words belonging to a term (in Table 2.2, this is called *gap*).
- **Order** - Terms must be in the same order as in the lexicon 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 lexicon are considered a lexicon 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 lexicon, 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 Table 2.2.

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.

	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

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 lexicon-based NER tools, finding that its performance in identifying terms from lexicons 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 lexicons in a number of formats and easily annotate texts.

2.5 Evaluation Metrics

For a certain task (for example, annotation of a corpus with terms related to diseases) 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 identification of disorder mentions, it could be an annotation done by an human expert. To calculate this measures we also need the number of true positives, 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;

[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`

- False positive (FP) – The system annotated a term that is not annotated in the reference;

- False negative (FN) – The system did not 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 ten terms annotated by the system, only six are annotated by the reference, then the system has a precision of 0.6. If every term identified by the system is also identified 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 it 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 eight terms of the ten 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.

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

2.5.1 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 MRRAD (Multilingual Radiology Research Articles Dataset) Corpus

To the best of my knowledge there is no parallel corpus of Radiology reports. So I created a Portuguese-English parallel corpus of research articles related to Radiology, assuming that the writing style and content of these research articles are similar to Radiology reports. For each research article the MRRAD corpus contains:

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 have constructed the corpus.

3.1.1 Web Crawl of the articles (1,2)

To obtain a list of research articles related to Radiology, that were available both in English and in Portuguese, I used 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 Topic] AND hasabstract[text]" (search done on Dec 11, 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[2] and only articles hosted in there were included in the corpus. More, only articles for which the original language is Portuguese are included in the corpus.

Finally, I programmatically crawled the SciELO pages for each article to get both the original Portuguese texts and the corresponding English translations. From the HTML of each page I extracted everything from the abstract until, but not including, the references/bibliography.

Three of the articles were surveys, not containing much vocabulary about Radiology (PMIDs: 19936506, 22002140, 23515770). They were excluded from the corpus. Other two contained encoding problems and were also excluded (PMIDs: 21793046 and 24263777).

The final result is a parallel corpus of 51 articles, distributed by journal as shown in Table 3.1.

To give a sense of the corpus size, the human English translations have a total of 163,423 words[3] the longer article having 12,451 and the smaller 848. The articles have an average of 3,204 words each.

3.1.2 Note On Human Translations

It is not known for sure how exactly the original human translations were performed, since some of the articles are not recent and some of the journals did not answer my request for more information about the translation, but all the

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

[2] http://www.scielo.br/

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

Table 3.1: Number of articles by journal in parallel corpus

Journal	Number Of Articles
Arquivos Brasileiros de Cardiologia	24
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

answers received mentioned the use of specialized translation services. Having said this, it is being assumed that the translations are of high quality since they are published by scientific magazines.

3.1.3 Yandex Translation (3)

I used Yandex's free Translate API[1] to machine translate the Portuguese version of the articles. Yandex is a Russian company which, among other things, sells automatic translation services, but it has a limited free service. It currently uses a Statistical approach to Machine Translation. Each translation request had a limit of 10,000 characters so I developed software 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]. I obtained Google's Statistical Machine translation using the *mt_translation* endpoint of the API and Unbabel's Machine Translation + Post-Editing using the *translation* API's endpoint. The requests for Unbabel Translations have a limit of words, so I used a software similar to the one utilized for the Yandex Translations.

[1] `https://tech.yandex.com/translate/`
[2] `http://developers.unbabel.com/`

3.2 Annotation

All the English versions of the articles in the corpus were annotated three times 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 am calling the three approaches Direct Match[1], All Match and Best Match[2]. Three different kinds of approaches were used to check what effect the annotation strategy have on the results, if any.

3.2.1 Direct Match - Annotation with Open Biomedical Annotator

The articles were annotated with OBA using the REST API[3]. The default parameters were used, namely the ones shown in Table 3.2.

Table 3.2: OBA parameters used

Parameter	Value
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 its 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.

[1]See 2.1.1
[2]See 2.4.3
[3]http://data.bioontology.org/documentation#nav_annotator

More information on how NOBLE Coder was used can be found at the MR-RAD GitHub repository[1].

3.3 Evaluation

For each document and annotation approach I created the set of the RadLex terms (identified by their RIDs) that were found in that document with that annotation approach. This is the data used in the assessment of translation solutions that follows.

The RadLex terms identified in each MT or MT+PE translated article were compared against the ones identified in the corresponding HT translated article, which was considered the 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. Consider that we have one Portuguese document and corresponding HT English translation and MT English translation. Four terms of interest were identified in the HT translation, bone, cell, finger, colon[2]. This is going to be our gold standard. In the MT translation, two terms of interest were found, brain, bone. One of these terms is also in the gold standard, which means TP = 1, but the other term is not, FP = 1. In the gold standard there are three terms that were not found in the MT translation, which means FN = 3. After calculations (see 2.5), 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.

[1] `https://github.com/lasigeBioTM/MRRAD/blob/master/notes_on_dataset_creation/using_noble_coder.md`

[2] I use here human understandable names instead of RIDs so that the example is easier to follow

Chapter 4

Experimental Results

4.1 NER Lexicon-based approach

Table 4.1: Number of RadLex terms found by document

Translation	Direct Match	All Match	Best Match
Human	119.55	177.92	145.0
Yandex	116.06	173.92	145.16
Google	120.8	179.49	147.61
Unbabel	120.92	178.86	148.16

Table 4.1 presents the number of RadLex identified by document using the different annotation approaches. 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 was expected since the All Match approach its the most flexible one in what it considers to be a mention of a RadLex term. The Best Match approach is more strict than the All Match approach but less than the Direct Match approach, considering lexical variations and word reordering, for example. But in all cases we can see that many terms are being identified in each document.

As seen in Figures 4.1, 4.2 and 4.3, the terms identified in Google translations are more similar to the ones identified in HT translations than the ones from Yandex translations. This could be just because the human translators used

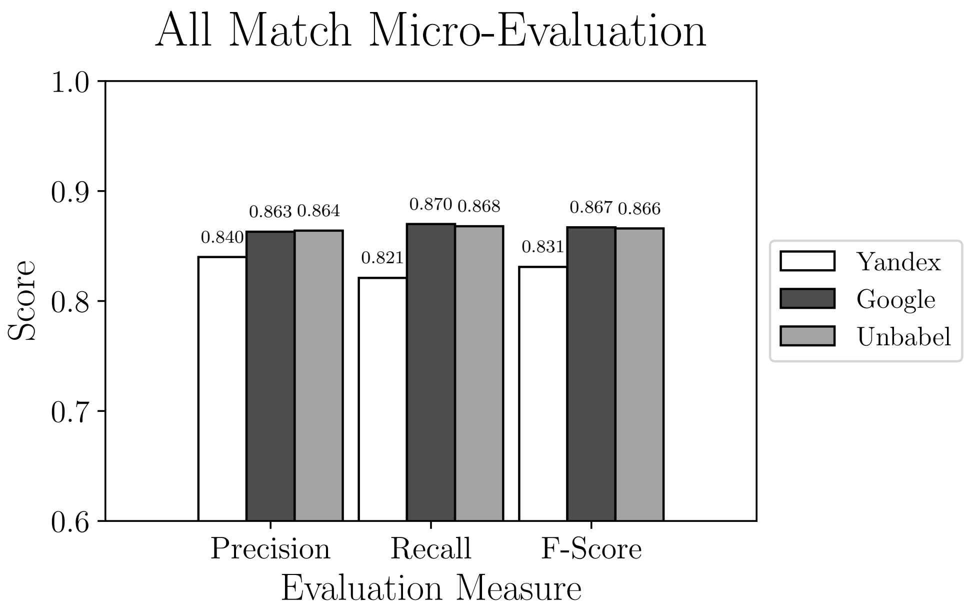

Figure 4.1: Micro Evaluation of Translations being tested (All Match)

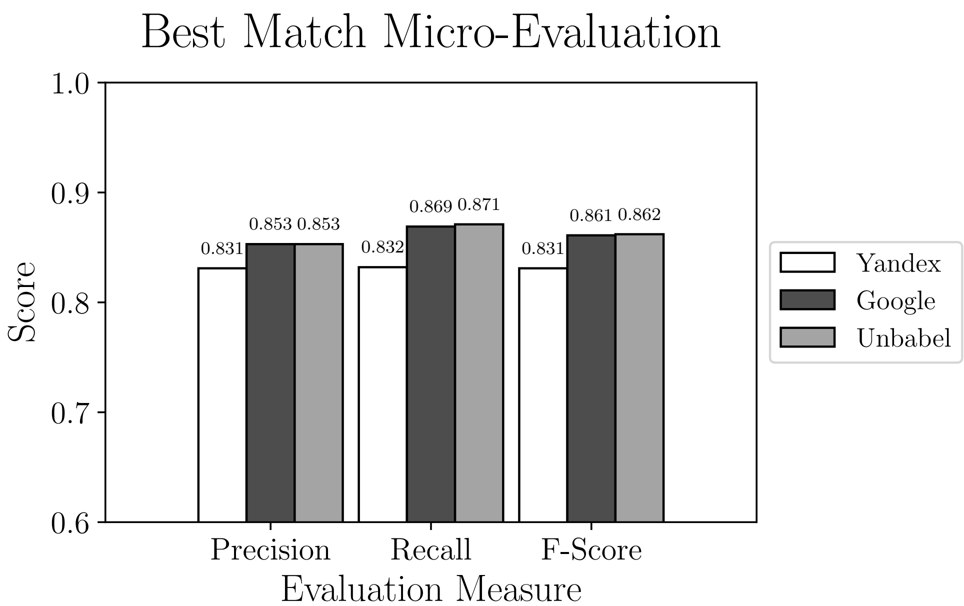

Figure 4.2: Micro Evaluation of Translations being tested (Best Match)

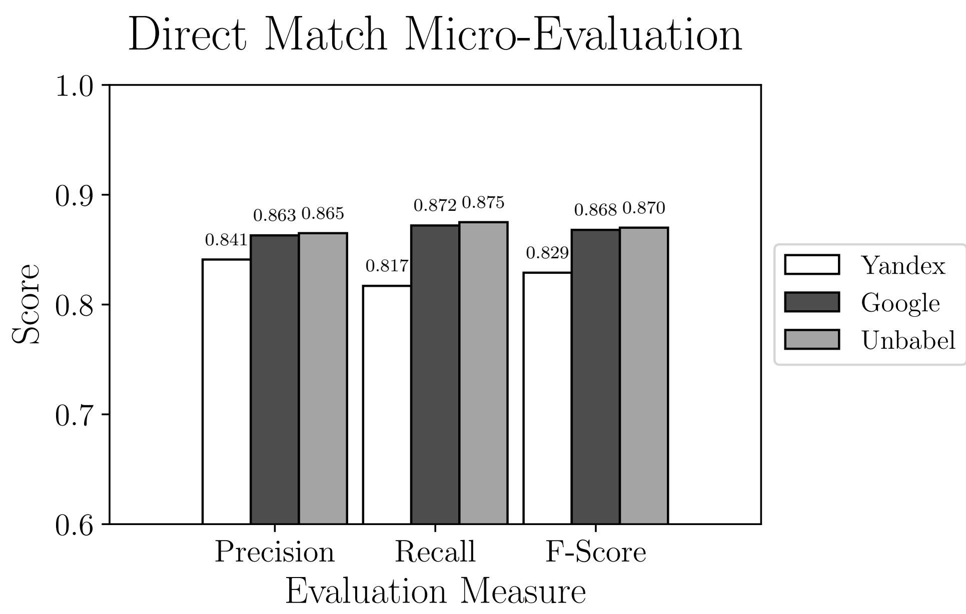

Figure 4.3: Micro Evaluation of Translations being tested (Direct Match)

Google Translator to help them in their translation process. This argument loses strength if we assume Google Translate translation outputs changed since the articles were human translated (publication years of the articles in MRRAD range from 2003 to 2013), but data could not be found to corroborate this assumption.

The terms identified in Unbabel and Google translations are really similar, the F-Scores being almost equal. That the translations are similar is not too surprising since the Post-Editing phase at Unbabel is done after MT translation using Google. What could be surprising is that Unbabel does not have a higher score. One conclusion to take from this is that Post-Editing step on the MT+PE does not add value for this task. The results are similar when a Macro Evaluation is done (see Appendix Tables 1, 2 and 3).

In the Introduction to the thesis I have proposed the following hypothesis:

Hypothesis: MT+PE is a good trade-off between quality and cost, compared with MT and HT, for translating Portuguese Radiology reports to English, for the purpose of identifying RadLex terms in the translated text.

I have written that for this to be true, "The terms identified in MT+PE translations have to be more similar to the ones identified in the HT translation than the ones identified in MT translations". This does not hold. So, for this task, if someone had to choose between Google and Unbabel, this someone would be better off using Google since it is cheaper.

I have also written that for the hypothesis to be true, "The terms identified in the MT+PE translations have to be similar enough to the ones identified in the HT translation". The terms identified in any of these translations are not extremely different but they are also not equivalent to the ones identified in the human translation. It could be the case that for some applications only translations close to human quality are acceptable, while for other applications a mediocre translation would be good enough. Therefore, the suitability of the MT and MT+PE translations probably depends on the practical usage for these translations and annotations.

To better understand the results I will now provide a detailed analysis on the annotations for the *clinical finding* and *anatomical entity* subtrees of RadLex. These are two of the subtrees that probably would be more important when applying RadLex to a Information Retrieval system, a type of application for which the results of this study can be useful.

4.1.1 Clinical Finding and Anatomical Entity Subtrees

Depending on the type of annotation approach and translation it was found between 35.25 and 55.55 *clinical finding* or *anatomical entity* terms per document (See Appendix Table 1). As seen in Figures 4.4, 4.5 and 4.6, the scores obtained are similar to the ones obtained for all terms, with Yandex translation identified terms being the less similar to the HT translation identified terms and Google and Unbabel having similar scores. Similar results were found when Macro evaluation was performed (see Appendix Figures 4, 5 and 6). But why these scores?

In an attempt to better understand the results, I did an analysis of the False Positives and False Negatives errors committed by the MT and MT+PE translations, focusing on the terms belonging to the *clinical finding* or *anatomical entity* RadLex subtrees. From preliminary analysis I knew that some of the FPs and

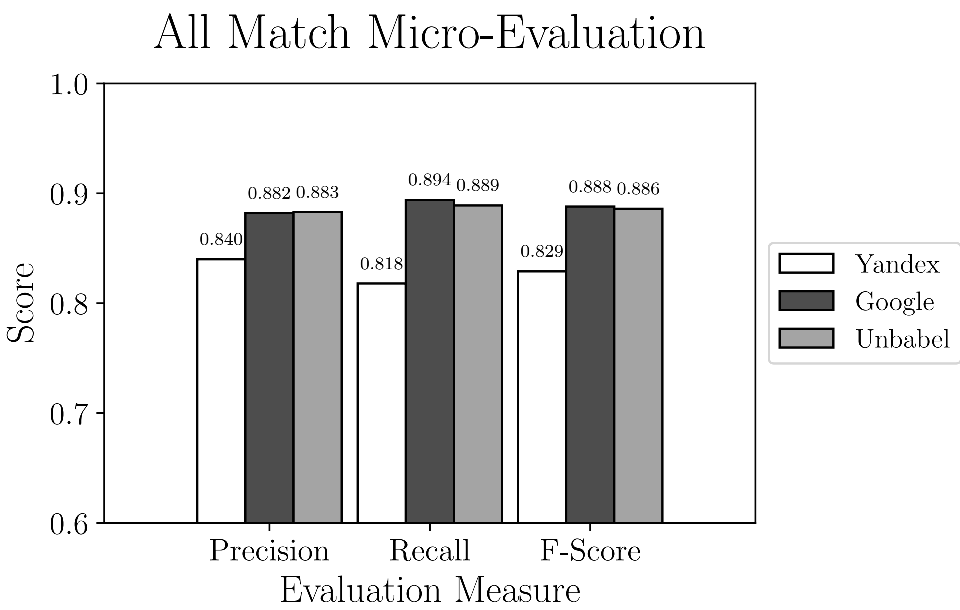

Figure 4.4: Micro Evaluation of translations being tested, considering just the terms from RadLex *clinical finding* and *anatomical entity* subtrees (All Match)

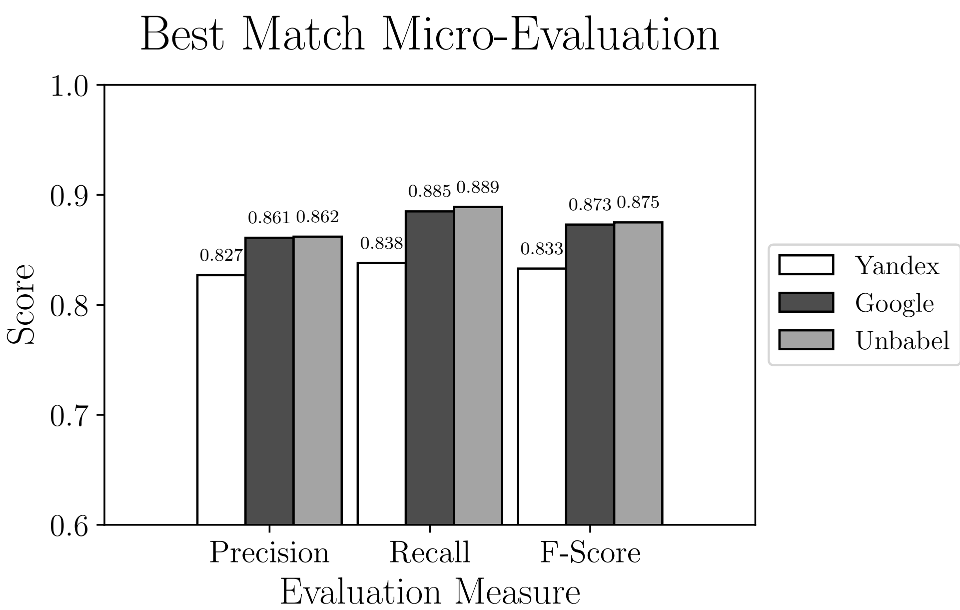

Figure 4.5: Micro Evaluation of translations being tested, considering just the terms from RadLex *clinical finding* and *anatomical entity* subtrees (Best Match)

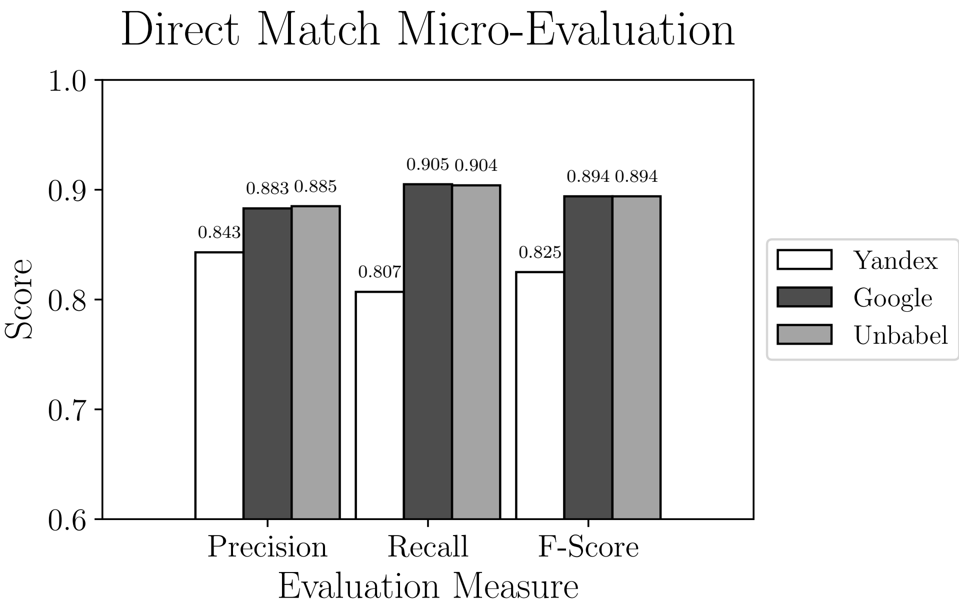

Figure 4.6: Micro Evaluation of translations being tested, considering just the terms from RadLex *clinical finding* and *anatomical entity* subtrees (Direct Match)

FNs are not caused by a erroneous translation but due to other causes, for example, an alternative translation which is correct but causes a different annotation, e.g., translating *parênquima pulmonar* to *pulmonary parenchyma* instead of to *lung parenchyma*. Both translations are correct but the second one leads to the identification of the term *lung* while the first does not. Still, I expected a higher number of real translation errors using Yandex compared with the Unbabel or Google translations, since both of these types of translation had better scores.

I did an analysis on the FPs and FNs errors committed by Yandex, Google and Unbabel translations in 9 random documents and each error was classified by type (See Supplementary Table 6). The results from the Best Match Approach were used. As predicted, the percentage of errors of Yandex due to a wrong translation (25% of 100 FPs or FNs) was higher than the percentage of errors of Google and Unbabel (22.09% of 86 and 21.18% of 85 FPs or FNs, correspondingly), but only slightly (See Appendix Table 2). The reasons for the others FPs and FNs included, among others, cases i) of different translations which are both correct but lead to different annotations, as described above and ii) in which the word identified does not have the same meaning in the text as it has in RadLex. For example, the case of identifying the anatomical term *hand* from "(...) on the other hand, it has to be considered that (...)" , in which the word *hand* is used metaphorically. This happens because a rule-based approach is being used, which does not consider the context of the term.

There were a lot of these ii) cases, which maybe would happen less if a Machine Learning NER approach was used. I thought about this but the problem is that, to the best of my knowledge, there is no data readily available to conduct an experiment of this type, i.e., I could not find English Radiology text resulting from human translation of Portuguese text and annotated with RadLex terms by experts.

Next I analyzed what kind of real translation errors were causing the FPs and FNs (See Supplementary Table 9). These subcategories included cases in which:

- There was an extra word in the translation;
- There was a missing word in the translation;
- A wrong hyphenation was used;
- An acronym was not translated;
- The test translation used a term that was too general;
- A wrong lexical variation was used;
- The most correct medical term was not used;

Each of these cases had a low number of occurrences and so it is not worth a deeper analysis. One interesting thing to note is that in the Yandex translations there were some cases (six) in which the original Portuguese word was not even translated. This never happened in the Google and Unbabel translations that were analyzed. This could be explained by the fact that probably Yandex focuses on different languages than Google and so their Portuguese-English translation and/or language models are not so well trained. But most of the errors correspond to just to a general wrong choice of terms to use as a translation. For example, translating *média* to *middle* instead of *mean* or *lesões de via biliar* to *lesions via bile* instead of *lesions to the biliary tract*. This type of problems could probably be solved by training Google and Yandex models with more data, specifically data related to medicine.

One could expect that Unbabel translations would have a lot less mistakes than Google's but this is not always the case. There are situations where errors are even added during the Post-Editing step. A review of the errors makes me propose that this could be due to the lack of medical knowledge of Unbabel current editors. For example, a stroke is something that occurs in the brain but in one case it was used as something that happens in the heart - someone with some knowledge on medicine would not make this error. But the truth is that Unbabel currently do not have a focus on medical content. I predict that if they did and invested in growing a crowd of experts with a better knowledge of medical language, this would lead to better results.

Chapter 5

Conclusions

In the Introduction of this thesis I wrote that I was going to answer the question "lacking the resources to pay for human translation services, what kind of automatic (MT) or semi-automatic translation (MT+PE) approach should be used in the task of translating Portuguese Radiology-related text to English, for the purposes of finding RadLex terms in the translated text?".

For this purposes, I have created the MRRAD corpus, a corpus of 51 Portuguese research articles related to radiology and four alternative translations to English for each one of these articles. This corpus can be used to study the efficacy of translation solutions in biomedical text, particularly text related to Radiology. To the best of my knowledge this is the first corpus of this type. This corpus could even be extended by other researchers, using different types of translation or languages, for example.

Using this corpus I did a quantitative evaluation of the performance of multiple automatic or semi-automatic translation approaches in the task of translating Portuguese Radiology-related text to English, for the purposes of recognizing RadLex terms in the translated text. To better understand the results I also did a qualitative analysis of the type of errors found. The results will certainly be helpful for the decision-making of developers who want to develop multilingual applications that apply Text Mining tools, specially in Radiology text. The results corroborates the conclusion that if the developers have limited financial resources to pay for Human Translations, they will be better of using a Machine Translation service like Google instead of a service that implements Post-Editing, like Unbabel. Of course, maybe there are better Machine Translation services than Google or better Machine Translation + Post-Editing services than Unbabel is currently offering for the medical field, and this is something that could be explored in further work.

Since this work explores a way to annotate non-English text using English terms, these results can motivate the sharing of annotations of biomedical text across communities. Linked-data (Barros & Couto, 2016) approaches, for example, will benefit from this sharing because they will have access to data that would be hard to access behind language barriers, which creates the possibility of developing semantic knowledge bases (Monteiro *et al.*, 2016) with multilingual content. This sharing will allow, for example, find reports from different languages when searching for Radiology reports about left shoulders.

In this dissertation I just assessed the application of recognizing RadLex terms from translated text. A more realistic approach would be to test the performance of each kind of translation in a real application, like a Information Retrieval (Manning *et al.*, 2009a) or Question Answering system. But even if we discover which translation strategy is better for each kind of system, the question of the feasibility of integrating translation in systems used in real-word settings (e.g. hospitals) remains and this is something that could be explored in further work, through, for example, a partnership with a clinical facility.

Doing translations of Radiology reports to be consumed by software its just part of what needs to done to break language barriers in this field. Web platforms like Radiopaedia[1], MyPACS[2] and AuntMinnie[3] have the goal of sharing radiological information in the Radiology community, but the information available is in English, which could be a obstacle to some radiologists. Not just because of difficulties in writing or reading English, but the fact that the text is not in the native language of the user can make her feel less welcome to the community. My point being that further work could explore the task of translating Radiology text for human consumption.

[1] `https://radiopaedia.org/`
[2] `https://www.mypacs.net`
[3] `http://www.auntminnie.com/`

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. 20

Ananda-Rajah, M.R., Martinez, D., Slavin, M.A., Cavedon, L., Dooley, M., Cheng, A. & Thursky, K.A. (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**, 1–8. 10

Antony, B.J. & Suryanarayanan Mahalakshmi, G. (2015). Content-based Information Retrieval by Named Entity Recognition and Verb Semantic Role Labelling. *Journal of Universal Computer Science*, **21**, 1830–1848. 3, 7

Barros, M. & Couto, F.M. (2016). Knowledge Representation and Management: a Linked Data Perspective. *IMIA Yearbook*, 178–183. 44

Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *To appear: EMNLP-2016*. 12

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. 19

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. 20

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. 18

Branco, A. & Silva, J. (2003). Tokenization of Portuguese: resolving the hard cases. Tech. rep., University Of Lisbon. 9

Bretschneider, C., Oberkampf, H., Zillner, S., Bauer, B. & Hammon, M. (2014). Corpus-based Translation of Ontologies for Improved Multilingual Semantic Annotation. In *Proceedings of the Third Workshop on Semantic Web and Information Extraction*, 1–8. 2, 20

Castilla, C. (2007). *Instrumento de Investigação Clínico-Epidemiológica em Cardiologia Fundamentado no Processamento de Linguagem Natural*. Ph.D. thesis, Universidade de São Paulo, São Paulo. 2, 17, 19

Cohen, A.M. & Hersh, W.R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, **6**, 57–71. 7

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. 10, 20

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*, **29**, 1233–46. 10

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. 22

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. 9, 10

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. 19

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. 18

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. 25

Gerstmair, A., Daumke, P., Simon, K., Langer, M. & Kotter, E. (2012). Intelligent image retrieval based on radiology reports. *European Radiology*, **22**, 2750–2758. 3, 7, 22

Groza, T., Köhler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F.M., Baynam, G., Zankl, A. & Robinson, P.N. (2015). Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database (Oxford)*, **2015**. 7

Hassanpour, S. & Langlotz, C.P. (2016). Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, **66**, 29–39. 1, 10

Hong, Y. & Kahn, C.E. (2013). Content analysis of reporting templates and free-text radiology reports. *Journal of Digital Imaging*, **26**, 843–849. 10

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. 21

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. 7

Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A. & Mark, R.G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, **3**, 160035. 11

Jonquet, C., Shah, N.H. & Musen, M.A. (2009). The open biomedical annotator. *Summit on translational bioinformatics*, **2009**, 56–60. 8, 22

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. 18

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. 18

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. 18

Klein, J.S. (2013). A Look Back at 2012 and Plans for 2013. *RadioGraphics*, **33**, 1–2. 22

Koehn & Philipp (2010). *Statistical Machine Translation*. Cambridge University Press. 12

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. 15

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. 22

LANGLOTZ, C.P. (2006). RadLex: a new method for indexing online educational materials. *RadioGraphics*, **26**, 1595–1597. 21

LANGLOTZ, C.P. & CALDWELL, S.A. (2002). The completeness of existing lexicons for representing radiology report information. *Journal of Digital Imaging*, **15 Suppl 1**, 201–205. 21

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2009a). Evaluation in information retrieval. In *Introduction to Information Retrieval*, chap. 8. 44

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2009b). Stemming and lemmatization. In *Introduction to Information Retrieval*, chap. 2.2.4. 9

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2009c). Tokenization. In *Introduction to Information Retrieval*, chap. 2.2.1. 9

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. 18

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. 10

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. 21

Mohit, B. (2014). Named Entity Recognition. In I. Zitouni, ed., *Natural Language Processing of Semitic Languages*, chap. 7, 221–246, Springer. 8

Monteiro, E., Sernadela, P., Matos, S., Costa, C. & Oliveira, J.L. (2016). Semantic Knowledge Base Construction from Radiology Reports. *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, **5**, 345–352. 44

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)*. 18

Nagata, M. & Pedersen, T. (2005). Proceedings of the ACL Interactive Poster and Demonstration Sessions. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. 18

Patil, S. & Davies, P. (2014). Use of Google Translate in medical communication: evaluation of accuracy. *The BMJ*, **349**. 18

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, 7

Porter, M. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, **14**, 130–137. 9

Randhawa, G., Ferreyra, M., Ahmed, R., Ezzat, O. & Pottie, K. (2013). Using machine translation in clinical practice. *Canadian family physician Medecin de famille canadien*, **59**, 382–383. 18

Robinson, P.N. & Bauer, S. (2011). *Introduction to bio-ontologies*. CRC Press. 10

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. 17

Somers, H. (2003). *Computers and translation: a translator's guide*. John Benjamins Publishing Company. 15

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*, **12**, 471–484. 22

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. 8, 25

Tatsumi, M., Aikawa, T., Yamamoto, K. & Isahara, H. (2012). How Good Is Crowd Post-Editing? Its Potential and Limitations. In *The Proceedings of Association for Machine Translation in the Americas (AMTA)*, San Diego, CA. 15

Toral, A., Noguera, E., Llopis, F. & Munoz, R. (2005). Improving question answering using named entity recognition. In *Natural Language Processing and Information Systems, Proceedings*, vol. 3513, 181–191, Springer, Berlin, Heidelberg. 3, 7

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**. xix, 8, 23, 24, 25

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. 18

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. 22

WEI, Q., CHEN, T., XU, R., HE, Y. & GUI, L. (2016). Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database (Oxford)*, **2016**, 1–8. 7

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. 19

WOODS, R.W. & ENG, J. (2013). Evaluating the completeness of radlex in the chest radiography domain. *Academic Radiology*, **20**, 1329–1333. 21

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. 19

ZHANG, Y., XU, J., CHEN, H., WANG, J., WU, Y., PRAKASAM, M. & XU, H. (2016). Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. *Database (Oxford)*, **2016**, 1–10. 7

Macro Evaluations

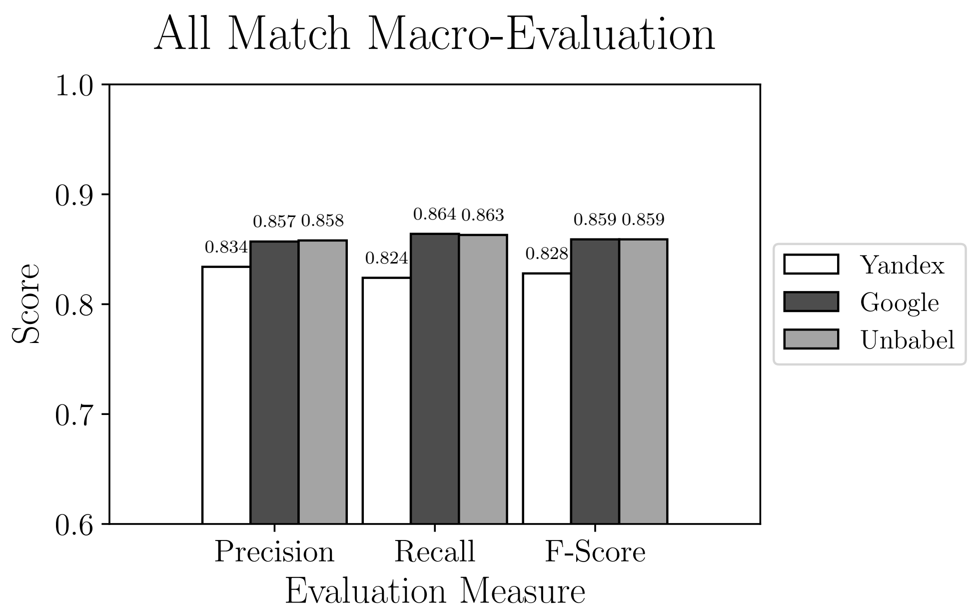

Figure 1: Macro Evaluation of Translations being tested (All Match)

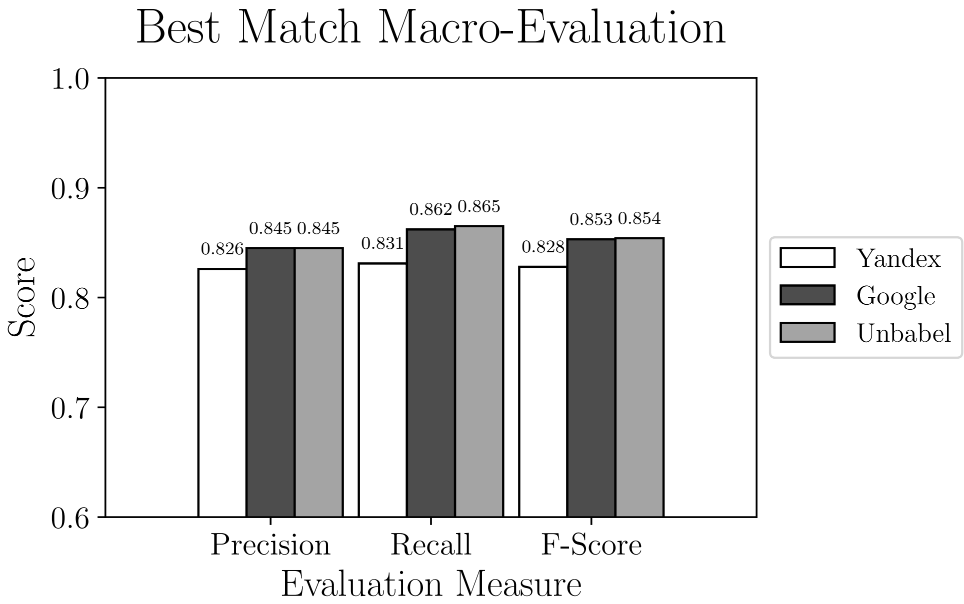

Figure 2: Macro Evaluation of Translations being tested (Best Match)

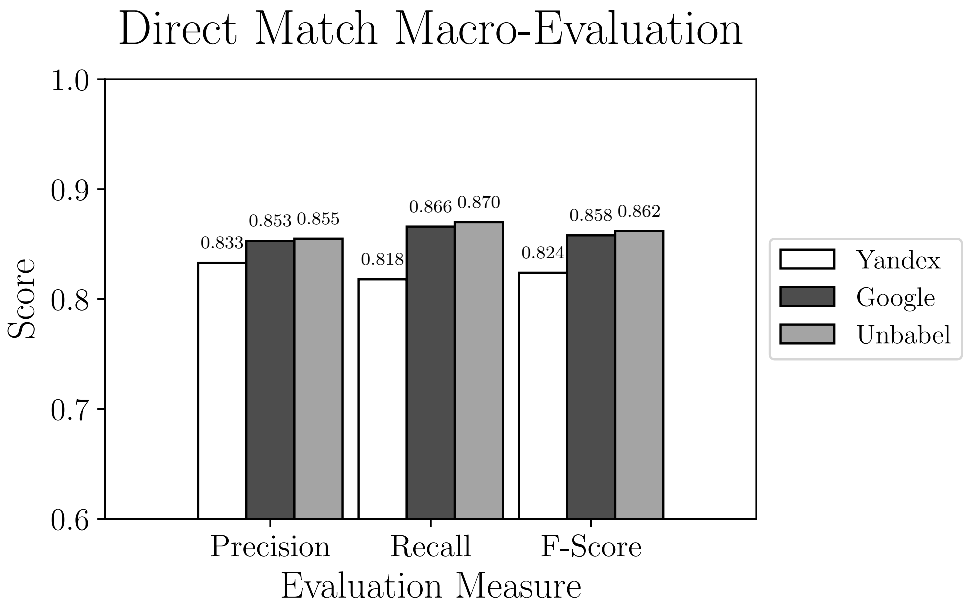

Figure 3: Macro Evaluation of Translations being tested (Direct Match)

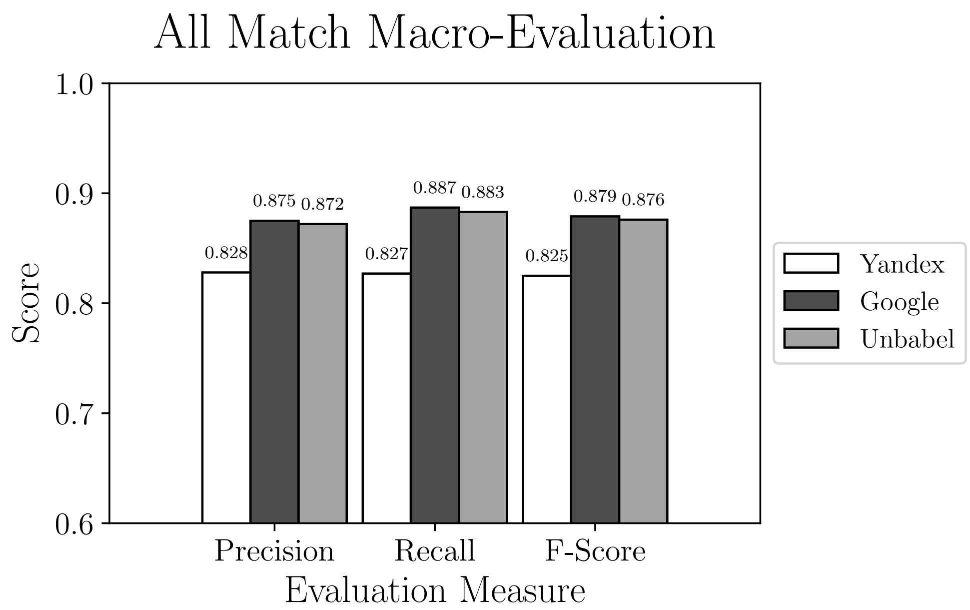

Figure 4: Macro Evaluation of translations being tested, considering just the terms from RadLex *clinical finding* and *anatomical entity* subtrees (All Match)

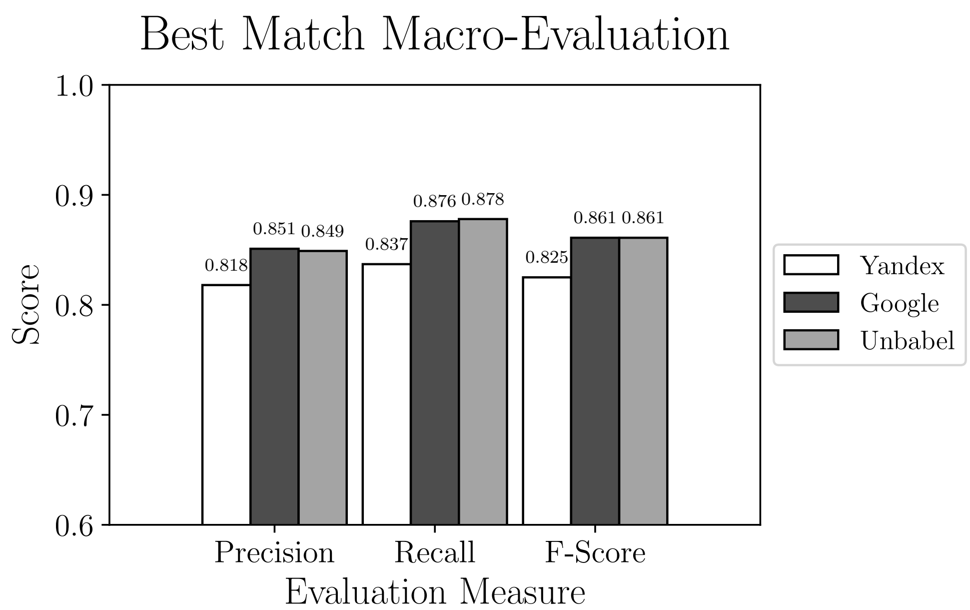

Figure 5: Macro Evaluation of translations being tested, considering just the terms from RadLex *clinical finding* and *anatomical entity* subtrees (Best Match)

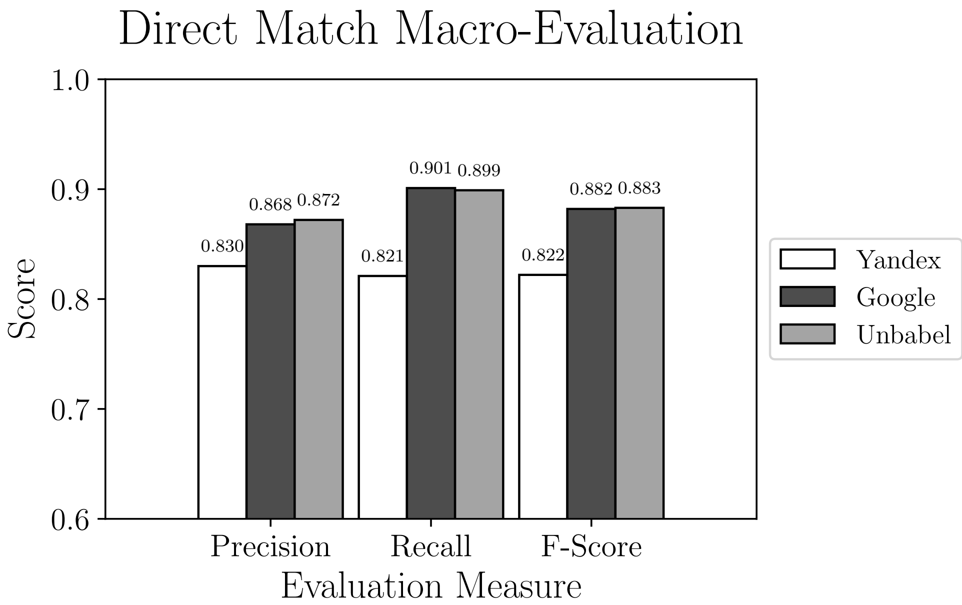

Figure 6: Macro Evaluation of translations being tested, considering just the terms from RadLex *clinical finding* and *anatomical entity* subtrees (Direct Match)

Subtrees Analysis

Table 1: Number of RadLex *clinical finding* or *anatomical entity* terms found by document

Translation	Direct Match	All Match	Best Match
Human	36.82	54.76	41.20
Yandex	35.25	53.33	41.75
Google	37.73	55.55	42.35
Unbabel	37.59	55.14	42.45

Error Evaluations

Table 2: Number of Errors Belonging to Each Category, by Translation Type. The column "?" contains the count of the errors for which I could not attribute a category

Translation	Wrong Translation	Not Wrong Translation	?	Total
Human	25	71	4	100
Yandex	19	64	3	86
Google	18	64	3	85
Unbabel	62	199	10	271