TRANSLATION OF UMLS ONTOLOGIES FROM EUROPEAN PORTUGUESE TO BRAZILIAN PORTUGUESE

Lucas Emanuel Silva e Oliveira[1], Sadid A. Hasan[2], Oladimeji Farri[2] e Claudia Maria Cabral Moro[1]

[1] Health Technology Post-Graduate Program, Polytechnic School, Pontifical Catholic University of Paraná. Curitiba, Paraná, Brazil
[2] Philips Research North America, Cambridge, Massachusetts, United States of America

Resumo: Ontologias terminológicas padronizadas e corretamente traduzidas são essenciais para o desenvolvimento de aplicações de processamento de linguagem natural na área da saúde. Para o desenvolvimento de uma aplicação de busca semântica em narrativas clínicas em português se fez necessária a utilização dos termos clínicos da Unified Medical Language System (UMLS). **Objetivos**: Traduzir termos da UMLS em Português Europeu para Português Brasileiro. **Métodos**: Foi desenvolvido um algoritmo de tradução semi-automática baseada em regras de substituição de texto. **Resultados**: Após execução do algoritmo e avaliação por parte de especialistas, o algoritmo deixou de traduzir corretamente apenas 0.1% dos termos da base de testes. **Conclusão**: A utilização do método proposto se mostrou efetivo na tradução dos termos da UMLS e pode auxiliar em posteriores adaptações de listagens em Português Europeu para Português Brasileiro.

Palavras-chave: Unified Medical Language System; Tradução; Processamento de Linguagem Natural.

***Abstract**: Correctly translated and standardized clinical ontologies are essential for development of Natural Language Processing application for the medical domain. To develop an ontology-driven semantic search application for Portuguese clinical notes we needed to implement the Unified Medical Language System (UMLS) ontologies, specifically for Brazilian Portuguese. **Objectives**: To translate UMLS terms from European Portuguese to Brazilian Portuguese. **Methods**: To develop a semi-automatic translation algorithm based on string replacement rules. **Results**: Following the experiments and specialists' evaluation the algorithm mis-translated only 0.1% of terms in our test set. **Conclusion**: The proposed method proved to be effective for UMLS clinical terms translation and can be useful for posterior adaption of a set of clinical terms from European Portuguese to Brazilian Portuguese.*

***Keywords**: Unified Medical Language System; Translation; Natural Language Processing.*

Introduction

Lexicon localization is often a determinant of the extent to which any natural language processing (NLP) application can be implemented or extended across multiple domains and languages. This is especially important for the clinical domain when standardized clinical ontologies and other gazetteers need to be leveraged for successful adaptation for other languages and regions. We developed an ontology-driven semantic search application for Portuguese clinical notes[1] exploiting UMLS clinical terms that includes a mixture of European Portuguese (pt) and Brazilian Portuguese (pt-br) terms. Adaptation of this application for pt-br requires an understanding of how certain words in pt differ orthographically or otherwise from their counterparts in Brazilian Portuguese and then using these insights to generate rule-based algorithms to translate the lexicon from one Portuguese dialect to the other.

During this study we did not identify any prior work that translates terms from pt to pt-br in the medical domain, but the work of the Natural Language Group at Systems Engineering and Computers Institute (INESC) shows us the main differences that we need to consider for this kind of translation. In their studies[2-4] they said: "The Portuguese from Portugal and Brazilian Portuguese differ in phonological, lexical, morphological and syntactic levels". The main objective of their work was to measure accurately the degree of difference between these two variants of Portuguese while translating different corpora (journalistic and technical). Some methods used in their work were replicated in Fernandes and Xatara dictionary translation[5], as well in this research, in addition to our main contribution which is addressing the peculiarities of terms in the medical domain (explained in the Methods Section).

We present a rule-based semi-automatic approach for translating UMLS terms originally documented in European Portuguese to corresponding terms in Brazilian Portuguese. Experimental results demonstrate the effectiveness of our approach.

Methods

We used the UMLS 2013AA Release, which contains 162496 absolute terms and 125817 unique terms from 4 ontologies, with terms written in both European Portuguese (pt) and Brazilian Portuguese (pt-br), as shown in Table 1.

Table 1 – Terminological lists of UMLS 2013AA Release

List	Language	Number of terms
Medical Dictionary for Regulatory Activities Terminology (MedDRA) Version 15.1, Portuguese Edition; MedDRA MSSO; September, 2012.	pt	92675
BIREME/PAHO/WHO;Descritores em Ciencias da Saude [Portuguese translation of Medical Subject Headings];Centro Latino-Americano e do Caribe de Informacão em Ciencias da Saude;2013;Sao Paulo (Brasil).	pt-br	65348
WHO Adverse Drug Reaction Terminology (WHOART). Portuguese Translation. Uppsala (Sweden): WHO Collaborating Centre for International Drug Monitoring, 1997.	pt	3750
The International Classification of Primary Care (ICPC). Portuguese Translation. Denmark: World Organisation of Family Doctors, 1993.	pt	723

As reference sources for the development of our translation algorithm, we used the ICD-10 (pt-br version – CID-10) ontology and a corpus of 8607 discharge summaries from multiple medical specialties from an academic medical institution in Brazil.

In the translation work by INESC[2-4], the researchers classified the types of contrasts (differences between pt and pt-br) by grammatical level (syntactic, morphological and lexical) and by usage frequency, where:

- Syntactic contrast: differs on text organization (sentence order, verb flexion, absence/presence of words, etc.).

- Morphological contrast: these include different derivation (prefixes and suffixes), different inflexion in the two variants or any morphologic alteration of the word, like different gender and number.
- Lexical contrast: words with orthographic or sense/connotation differences.
- Frequency of use: words not shared by the two variants or present great disparity of use.

As our objective is to translate ontology terms only and not texts with complete sentences, we will focus on lexical and morphological contrasts. As the syntactic contrast is not important in this work, we perform Word-level translation instead of using all the words of a term together.

Our methodology comprises the following steps:

Identify UMLS terms with Accurate Translation

The UMLS terms from the pt ontologies that were present in the discharge summaries and/or CID-10 were considered correct (i.e. we do not need to translate them as these terms are essentially the same in both European and Brazilian Portuguese).

- 4784 UMLS terms were found in discharge summaries
- 1396 UMLS terms were found in CID-10

These terms were manually checked by one of the authors, who is a native pt-br speaker and familiar with clinical terminology. Only one term, "Hiperkalemia", was incorrect. This term was incorrectly spelled in one discharge summary and its correct form in pt-br is Hipercaliemia or Hiperpotassemia.

All correct UMLS terms found were removed from our consideration for further processing. Note that, some terms occurred both in discharge summaries and CID-10, so in total we have 5619 unique terms that were marked as correct and removed. We used the remaining 120198 UMLS terms in the next steps of our translation algorithm.

Word-level analysis

All other terms not identified in the previous step as accurate were tokenized into constituent words (e.g.: [Insuficiência cardíaca] → [Insuficiência], [cardíaca]), and the frequency of each word was computed.

There were 58123 words (a sample is shown in table 2), and the words with more than 4 occurrences in the UMLS were manually verified by the same researchers that checked the terms in step one, resulting in the top 9701 words (16.7%) selected as our training set.

Table 2 – Partial frequency table of UMLS words

Word	Number of occurrences
de	31361
síndrome	1780
células	1262
vírus	917
congénita	537
injecção	132

After verifying the top 9701 words looking for those incompatible with pt-br words, we found that 166 words (1.71% of top words list) were contrasting (amounting to 6168 occurrences). The remaining 48422 words were used as our testing set.

We investigated the contrasts list (i.e. the words that are written differently in pt and pt-br) and observed some similarities in most of the words (93 out of 166 words). These similarities are similar to that were found in INESC research[2-4], where they extracted automatically some orthographic errors using some string sequences. This motivated us to build a set of simple string pattern replacement rules to automatically translate these words from pt to pt-br. The rules and word examples are shown in Table 3. The rules used in the INESC work that were not found in any contrast occurrence were removed from our rule set.

Table 3 – String replacement rules and word examples

String to Replace	Replaced by String	Original word	Replaced word
act	at	Fractura	Fratura
		Actividade	Atividade
ect	et	Rectal	Retal
		Afectivo	Afetivo
oct	ot	Nocturno	Noturno
uct	ut	Fructose	Frutose
opt	ot	Óptico	Ótico
		Adoptiva	Adotiva
pç	ç	Adopção	Adoção
mn	n	Polisomnografia	Polisonografia
cç	ç	Reacção	Reação
		Injecção	Injeção
cc	c	Direccional	Direcional
		Seleccionado	Selecionado
gén	gên	Congénita	Congênita
		Oxigénio	Oxigênio

Furthermore, all words with erroneous accentuation that are not covered by “gén to gên” rule (38 out of 166 words) were included in another set of replacement rules, presented in Table 4.

Table 4 – Accentuation replacement rules and word examples

Character to Replace	Replaced by Character	Original word	Replaced word
ó	Ô	Insónia	Insônia
		Isotónica	Isotônica
		Crónicas	Crônicas
		Económico	Econômico
é	Ê	Bebé	Bebê
		Epidémico	Epidêmico
		Esquizofrénica	Esquizofrênica

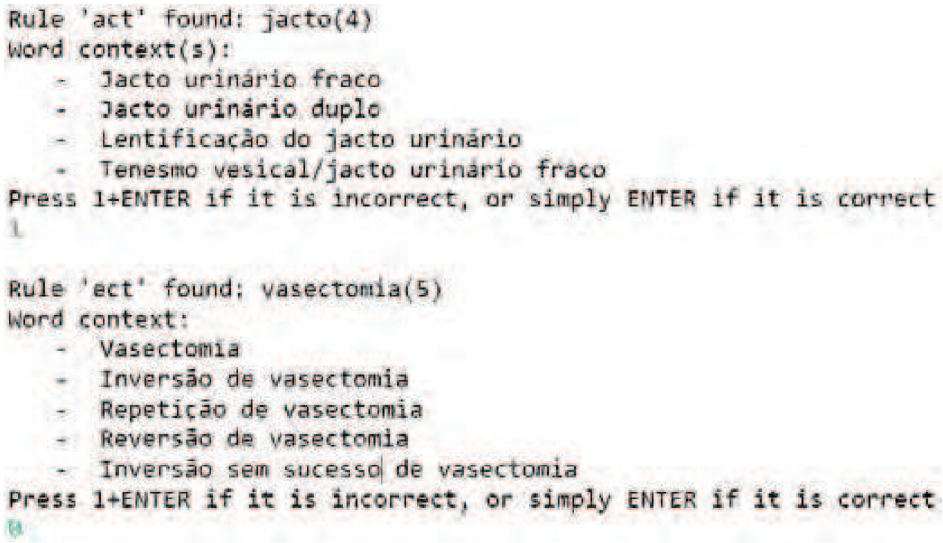
While automatically translating the test set words (the remaining 48422) to pt-br using this rule-based approach, it was realized that they did not always correctly replace the words, thereby generating several translation errors. For example, there are some words where the "act → at" rule causes a misspelling error, like: "Fusobacteria", "Galactose", "Lactentes", etc. Similar errors occur with other replacement rules as well.

Also, we encountered some morphological contrasts where it is necessary to look at the context of the entire UMLS term, not just the single word, to assign the right "gender" to the phrase e.g. when the words "Hormona" (Hormone) and "Tireoideia" (Thyroid) co-occur. The word "Hormona" when translated to pt-br changes the "gender" of the word. To address this change, we had to also transform its contextual words. For instance, the pt term "Anomalias das hormonas sexuais masculinas" adapted to pt-br has to be "Anomalias dos hormônios sexuais masculinos". To resolve this problem we had to revise the previous preposition and the subsequent adjective's gender.

The word "Tireoideia" may be difficult to translate depending on the context. For example, the pt term "função tiroideia anormal" translated to pt-br has to be "função tireoidiana normal", and "Cancro anaplásico da tiroideia" has to be "Cancro anaplásico da tireóide". Some specific rules were defined to translate such cases. We discovered that when the word "tiroideia" comes after a preposition (na, pela, da, etc.), we have to translate the word to "tireóide". For other scenarios, we have to define the translation based on the gender and the number of previous words to make the decision between "tireoidiana", "tireoidianas", "tireoidiano" and "tireoidianos".

Semi-automatic rule-based translation interface

By considering the scenarios presented above, we developed a semi-automatic translation approach to consider the context of words to determine the rule(s) that is/are the best match. For this purpose, we built an interface to display the occurrences found and automatically generate a word substitution command if the user marks the word as a contrast. A snapshot of the command-line interface is shown on Figure 1.



```
Rule 'act' found: jacto(4)
Word context(s):
    -  Jacto urinário fraco
    -  Jacto urinário duplo
    -  Lentificação do jacto urinário
    -  Tenesmo vesical/jacto urinário fraco
Press 1+ENTER if it is incorrect, or simply ENTER if it is correct
1

Rule 'ect' found: vasectomia(5)
Word context:
    -  Vasectomia
    -  Inversão de vasectomia
    -  Repetição de vasectomia
    -  Reversão de vasectomia
    -  Inversão sem sucesso de vasectomia
Press 1+ENTER if it is incorrect, or simply ENTER if it is correct
```

Figure 1 – Snapshot of semi-automatic rule-based translation interface

Results and Discussion

After running the translation algorithm, 7442 (5.9%) of 125817 UMLS terms were translated. This value is similar to the INESC group's conclusion that "generically there is a global 10% discrepancy between the two variants"[2]. The 4% difference can be explained by the fact that the UMLS corpus already has a mixture of pt and pt-br.

To evaluate the translation accuracy, a test set with 10000 randomly selected UMLS terms (almost 8% of total UMLS terms) was defined. The test set maintained the balance between translated (590 – 5.9%) and non-translated terms (9410 – 94.1%). Two specialists validated the test set and found that only 11 were not translated correctly (0.11%). One specialist was a nurse and master's student with significant experience in clinical care, and the other was a 4th year medical student, both fluent in pt-br.

After analyzing these terms, we found some reasons behind their incorrect translation. Two terms do not have accentuation in pt ("porfiria cutanea" and "calendario de imunizações"), but have in pt-br ("porfiria cutânea" and "calendário de imunizações").

One term has a different word in pt (Atrofia secundária difusa da coroideia) than pt-br (Atrofia secundária difusa da coróide), somewhat similar to the "tiroideia" issue.

The term "Reticulossaarcoma com compromisso dos gânglios linfáticos intratorácicos" was incorrectly written originally in the UMLS pt list. The correct term should be "Reticulossarcoma com compromisso dos gânglios linfáticos intratorácicos".

And finally, seven terms had the word "quisto" that means the same as "cisto" in pt-br, but the specialists marked it as wrong because "cisto" is more widely used, making these terms wrong by their frequency of use.

If we consider the fact that one of these errors was actually caused by a misspelling in the original UMLS list, we had 10 incorrect terms in the test set, leading us to believe that if in 8% of terms we had 0.1% incorrect terms, this value tends to remain close to that for the entire list of UMLS, showing the effectiveness of our rule-based algorithm.

The word-level analysis seems to be the best way to translate an entire ontology, since with only single-word correction we achieved the translation of a lot of terms simultaneously.

Finally, some words are completely different; we could not do automatic translations in such cases. And it is worth noting that after the last orthographic agreement between Portuguese-speaking countries[6, 7], the differences between pt and pt-br have decreased, easing the amount of work on this kind of translation.

Conclusion

The main differences between European and Brazilian Portuguese are orthographic, and can be identified by a simple set of substrings, which allowed us to use a rule-based approach to translate the UMLS clinical terms. However the rules can correct some terms and at the same time harm others, thus necessitating the implementation of semi-automatic translation.

The word-level analysis proved to be effective on multiple terms' simultaneous translation, reducing the amount of work for analyzing the UMLS terms.

The proposed method can be reused on other ontologies that need translation from European Portuguese to Brazilian Portuguese.

References

[1] Hasan SA, Zhu X, Liu J, Barra CM, Oliveira L, Farri O. Ontology-Driven Semantic Search for Brazilian Portuguese Clinical Notes, Proceedings of the 15th World Congress on Health and Biomedical Informatics, Sao Paulo, Brazil,19-23 August 2015.

[2] Wittmann LH, Pereira M de J. Português Europeu e Português Brasileiro: alguns contrastes. Actas do X Encontro Nacional da Associação Portuguesa de Linguística; 1994. p. 613.

[3] Wittmann LH, Pêgo TR, Santos D. Português Brasileiro e Português de Portugal: algumas observações. Actas do XI Encontro Nacional da Associação Portuguesa de Linguística. 1995;2-4.

[4] Barreiro A, Wittmann LH, Pereira M de J. Lexical differences between European and Brazilian Portuguese. J Res Dev [Internet]. 1996;5(2):75–101. Available from: http://www.linguateca.pt/Repositorio/Barreiroetal95.pdf.

[5] Fernandes, HYS; Xatara, CM. Kernerman French-Portuguese dictionary: adapting the translation from European Portuguese to Brazilian Portuguese. Kernerman Dictionary News, n. 19, p. 6-9, 2011. Available from: http://hdl.handle.net/11449/122764.

[6] BRASIL. Decreto nº 6.583, de 29 de setembro de 2008. Promulga o Acordo Ortográfico da Língua Portuguesa, assinado em Lisboa, em 16 de dezembro de 1990. Diário Oficial, Brasília, DF, 29 set. 2008. Seção 1, p. 1.

[7] BECHARA, Evanildo. A nova ortografia. São Paulo: Nova Fronteira, 2008.

Contato

Lucas Emanuel Silva e Oliveira
PhD student in Health Informatics at
Health Technology Post-Graduate Program,
Polytechnic School, Pontifical Catholic
University of Paraná. Curitiba, Paraná, Brazil.
Professor at Polytechnic School, Pontifical
Catholic University of Paraná. Curitiba,
Paraná, Brazil.
Phone: +55 41 99287-2702
E-mail: lucas.oliveira@pucpr.br
Address: R. Imaculada Conceição, 1155 –
Rebouças, Curitiba/PR.