Machine Translation on the Medical Domain: The Role of BLEU/NIST and METEOR in a Controlled Vocabulary Setting

Andre CASTILLA **Alice BACIC** **Sergio FURUIE**

Informatics Service, Instituto do Coracao
University of Sao Paulo
Av. Dr. Eneas de Carvalho Aguiar 44
Sao Paulo, Brazil, 05403000

castilla@terra.com.br alice.bacic@incor.usp.br sergio.furuie@incor.usp.br

Abstract

The main objective of our project is to extract clinical information from thoracic radiology reports in Portuguese using Machine Translation (MT) and cross language information retrieval techniques. To accomplish this task we need to evaluate the involved machine translation system. Since human MT evaluation is costly and time consuming we opted to use automated methods.

We propose an evaluation methodology using NIST/BLEU and METEOR algorithms and a controlled medical vocabulary, the Unified Medical Language System (UMLS). A set of documents are generated and they are either machine translated or used as evaluation references. This methodology is used to evaluate the performance of our specialized Portuguese - English translation dictionary.

A significant improvement on evaluation scores after the dictionary incorporation into a commercial MT system is demonstrated.

The use of UMLS and automated MT evaluation techniques can help the development of applications on the medical domain. Our methodology can also be used on general MT research for evaluating and testing purposes.

1 Introduction

Machine Translation (MT) of medical texts has a potential role on catastrophe crisis management. International emergency teams are frequently mobilized in these situations. These teams are mostly multidisciplinary and involve professionals of different nationalities who possibly do not share the same language between themselves or the patients. Such situations of language diversity are a promising field for MT and Cross Language Information Retrieval (CLIR) applications. These techniques could certainly facilitate information exchanging within the professionals and between professionals and patients. We are currently researching CLIR as a tool for clinical information extraction from medical texts on the chest radiology domain. There are three basic approaches to CLIR, based on computational translation, on concepts and on parallel corpus (Oard, 1996). The main objective of our work is translating Portuguese chest radiology reports aiming clinical information extraction from queries in English.

An important feature of the MT system on this task is the correct manipulation of the terms and concepts of the specialized domain. Specialized texts have innumerable specific terms, many of them composites, the main goal is correctly identify and process them with high quality.

The most important and costly phase of this approach is the specialized translation dictionary elaboration. At this phase an instrument for probing our lexicon will be needed in order to evaluate its performance. Manual MT evaluation uses of human resources that classify performance according to subjective and objective criteria (Hovy, 2002). Such evaluations are expensive inasmuch they involve the use of onerous human resources. So we opt to use automated MT evaluation tools during the initial phases of the dictionary elaboration due to its low cost and high reproducibility.

We propose a methodology of automated MT evaluation of medical terms which uses the Unified Medical Language System (UMLS). The UMLS is a specialized knowledge base which contains a multilingual controlled vocabulary used here as a tool for the development of the dictionary. We have a hypothesis that a satisfactory performance on terms translation in a controlled dictionary can credence the system for use in texts of the same domain.

1.1 UMLS

The UMLS is a project of the National Library of Medicine of the National Institute of Health (Bethesda, USA) that integrates different sources of knowledge in a single database. It is composed of three parts. The *Metathesarus* is the central set that unifies several medical vocabularies and classifications in complex database. To each *Metathesarus* entry is assigned a string unique identifier (SUI). The semantically equal strings are mapped to only one concept which also has a unique identity. Thus we have a many-to-one relation between strings and concepts. Table 1 shows the entailed strings to the chewing gum concept (CUI –C0008037).

SUI	STR	LAT
S2402143	kauwgum	DUT
S2402142	kauwgom	DUT
S0024341	chewing gum	ENG
S0024341	chewing gum	ENG
S0363185	chewing gum	ENG
S0024342	chewing gums	ENG
S0046367	gum, chewing	ENG
S0046374	gums, chewing	ENG
S1858463	purukumi	FIN
S0229117	gomme a macher	FRE
S0275481	chewing gum	FRE
S0275482	pate a macher	FRE
S1508737	kaugummi	GER
S2083076	gomma da masticare	ITA
S0435564	goma de mascar	POR
S1111589	zhevatel'naia rezinka	RUS
S0453178	goma de mascar	SPA

Table 1- All strings mapped to the concept C0008037. This concept is member of the substance semantic type - SUI – string unique identifier, STR- string, LAT – language.

This concept grouping allows an inference that the strings diverse in languages can be considered equivalent and able to be used in translation. Studies had already used this property for quality improvement to the statistical MT system. (Eck, 2004)

Another important feature of the concept is the semantic type. This classifies the several concepts of the UMLS in 134 classes whose relations are described on *Semantic Network*, the second part of the UMLS. The *Semantic Network* specifies the potential relations between the concepts on a binary form. However, these associations are very generic and a manual analysis showed that only 17% are correct and 38% have some significant information (Vintar, 2003). Additionally, the UMLS preserves the original relations from each source through common structure.

The third part is the *Specialist Lexicon* which is a comprehensive set of English words, as well as its flexion rules, diverse information and the canonic forms. The sources of the words are the strings of *Metathesarus* as well as other consecrated English dictionaries.

1.2 MT Automated Evaluation

The automated evaluation of MT is carried through by using reference translations for matching the translated texts. There are the N-gram co-occurrence evaluations which analyze the agreement of terms and their sequences (N-gram) between the evaluated and reference texts. These group representatives are the BLEU algorithm and its derivative, the NIST algorithm (Papineni 2002; NIST 2001). They both calculate the accuracy of the translation comparing it with the reference translations and incorporate a size penalty. There is some correlation between BLEU scores and human quality judgments (Papineni 2002).

The results are obtained tabulating the N-gram fraction of the translation evaluation that also occurs in the reference translation. The BLEU Algorithm measures the quality as a weighed sum of the counts of co-occurred N-grams while NIST variant uses the geometric mean. Both algorithms include penalty for the translations whose sizes differ significantly from the reference translations; yet on NIST algorithm this was changed to diminish impact of the small variations. N-gram co-occurrence evaluations usesegments as units which, in our case, will be UMLS concepts. Each segment is scored and then the results are accumulated.

A new automated MT metric has been recently proposed. METEOR system automatically works by computing the unigram precision and the recall between the terms of the evaluated and reference translations. The evaluation also proceeds in sequence of stages. At the first stage all exact matches are detected between the two strings, while at the second stage the words not matched in the first one are stemmed using the Porter stemmer, and then matches are found between these stemmed words. It is more reliable than BLEU/NIST scoring at sentence level translations. The METEOR produces scores in the range of [0,1] based on a combination of unigram precision. There is a penalty related to the average length of matched segments between the evaluated translation and its reference (Jayaraman, 2005; Lavie, 2005).

2 Methodology

Our experiment evaluates the performance of the specialized translation dictionary. Each entry of the dictionary consists of a Portuguese word or expression, its English counterpart and the grammatical class. It was developed following a multi-stage development workflow designed for MT purposes. (Dillinger, 2001). The initial phase was the lexical objectives definition. The dictionary sources were then selected. The first words were extracted from thoracic radiology reports, which is the focus of our information retrieval project. The words in Portuguese had been manually translated into English. The second source was the specialized words extracted from abstracts of Radiology and Radiographics journals. These words were grouped into a dictionary developed for the word processor orthographic correction (Chang, 2003). These words in English were translated into Portuguese. The third origin was the Radlex, an initiative to elaborate a controlled lexicon for the Radiology domain. It's a descendant of the American College of Radiology Index of Radiological Diagnosis, and the released draft version contains only thoracic radiology terminology (RSNA, 2005). Finally, we selected words from terms of the Portuguese version of the Medical Subheadings (MESH), a specialized classification for medical literature. These words were translated into English. All these entries were incorporated in a spread sheet, classified and corrected, forming the basic dictionary. Then, the dictionary was incorporated into the MT system and the initial translations were carried through. A selection of dictionary entries is shown on table 2.

Portuguese	English	Category
broncopulmonar	bronchopulmonary	Adjective
histopatologicamente	histopathologically	Adverb
derrame pleural	pleural effusion	Subject
esclerosando	sclerosing	Verb

Table 2- Sample entries of the specialized dictionary.

Finally, on basis of translation adequacy, corrections and adjustments were made. The entries of our specialized dictionary are the full forms, including the number and gender inflections, as well as the splitting hyphenation variants compound terms. Despite the defined grammatical rules for the use of hyphen and the compound words creation, wide morphologic variation was observed in clinical texts. The entries can contain isolated words, compound words as well as bigger expressions. Preference was given to the use of isolated words when possible. We used lengthier expressions to improve translation performance when necessary.

The experiment began with selection of UMLS concepts to be processed. It was made in two ways, matching the lexical and thus semantic objectives of our domain. We selected only concepts which have a string in Portuguese.

The first way was choosing concepts from five semantic types. The second method was through the relationships contained in the UMLS. An algorithm searched for all child concepts from the thorax concept (CUI - C0039992), according to MESH relationships stored within UMLS.

The BLEU/NIST algorithms deliver absolute scores which could not be favorable for isolated evaluations. This is the main motivation for our methodology, to evaluate our dictionary creating a set of references from a controlled vocabulary. From the UMLS Metathesarus, we will take a set of strings mapped to the same concept. Some will be machine translated and evaluated, while others will function as references. Thus we can simulate upper and lower theoretical limits of MT performance.

We have five groups of concepts. Four of them are derived from the UMLS concept types, and the fifth is composed of the thorax child concepts (TCC). Each group was evaluated independently as one document. Each concept of this document will be a segment unit for MT evaluation. From one concept we picked four different mapped strings. The first is a Portuguese one and will be collected to the SOURCE group. This string will be machine translated and then evaluated. Secondly, we picked two distinct strings in English, one collected from the REF group (global reference) other from the HIGH (high limit score). Finally, we selected a term in a language other than English and Portuguese that will be collected to the LOW group. Each group will be treated as a document and all having the same collections of concepts grouped in different languages. Group SOURCE will be submitted to MT with and without the use of the specialized dictionary generating the groups TEST and DRY respectively. The groups TEST, HIGH, LOW and DRY will then be compared with the group REF through algorithm NIST/BLEU and METEOR, generating a set of four scores for each document. The graphical representation of our methodology is depicted in figure 1.

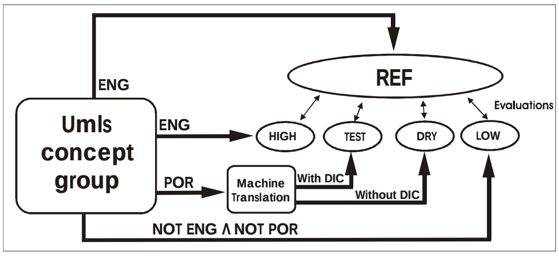

Figure 1- Graphical representation of the methodology. Each UMLS concept group generates 5 documents used on the experiment. The REF group is the global reference for scoring the other four documents (HIGH, TEST, DRY, LOW) by both algorithms.

The experiment was run on the version 2004AD of UMLS. The evaluations were carried through by script BLEU/NIST version 11 distributed by NIST and the algorithm METEOR version 0.4.3. The MT system used is the Systran Premium 4.0 (Systransoft, San Diego).

3 Results

The five semantic types selected to populate the test groups and the number of Portuguese strings available for processing in each group are shown on table 3.

GROUP	n
Acquired Abnormality (AA)	675
Body Location or Region (BLR)	57
Body Part, Organ or Organ Component (BPOC)	898
Body Space or Junction (BSJ)	71
Tissue(T)	94

Table 3- Selected semantic types and number of strings in Portuguese available for each group.

The sixth group was created searching for child relations to the thorax concept according to the MESH definitions within five levels. Of the 3240 selected concepts only 116 had strings in Portuguese which formed the last test group.

The set of NIST scores for each test group is listed on table 4.

GROUP	HIGH	TEST	DRY	LOW
AA	11,56	4,40	2,80	1,14
BLR	6,41	2,08	0,92	0,66
BPOC	11,13	3,50	1,84	1,09
BSJ	7,48	2,24	1,04	0,16
T	8,02	3,06	1,48	0,69
TCC	8,15	3,41	1,35	0,86

Table 4- Accumulated NIST scores for each tested document.

The set of METEOR scores for each test group is shown on table 5. The F-mean is a harmonic mean weighted more heavily on recall than precision.

	High		Test		Dry		Low	
GROUP	S	Fm	S	Fm	S	Fm	S	Fm
AA	0,95	1,00	0,40	0,56	0,14	0,23	0,10	0,17
BLR	0,99	1,00	0,50	0,61	0,29	0,38	0,08	0,14
BPOC	0,91	1,00	0,29	0,48	0,11	0,21	0,11	0,14
BSJ	0,95	1,00	0,37	0,52	0,17	0,27	0,13	0,19
T	0,93	1,00	0,42	0,62	0,17	0,27	0,09	0,15
TCC	0,96	1,00	0,36	0,52	0,12	0,20	0,05	0,07

Table 5- The Global score (S) and the F-mean(Fm) for each tested document.

4 Discussion

The elaboration of the dictionary follows a workflow that systemizes in multiple stages the acquisition after lexical objective definition. Our approach is to manually create a reusable knowledge base obtaining high quality MT through intensive specialist labor. To cover the terms related to thoracic radiology domain, we need a significant number of nouns and adjectives since clinical reports are mainly descriptive rather than narrative. Our specialized dictionary contains 2743 nouns, 86 adverbs, 44 acronyms, 2734 adjectives and just 33 verbs. The dictionaries terms were selected from specialized sources closely related to our project semantic objectives. A system of evaluation of lexical needs could assist the dictionary elaboration (Dillinger, 2001).

This evaluation would be ineffective if the processed concepts differed significantly from the domain of our project. Descriptive radiology reports can be summarized as a collection of anatomical structures associated to imaging features. So we decided to choose a restricted group of semantic types and to collect concepts related to our anatomical focus. Unfortunately the UMLS Semantic Network proved to be ineffective for related concept selection on our work. The network maps the possible relations between Metathesarus concepts, but these are broad and generic. By using UMLS Semantic Network, we obtained a big number of paired concepts on which manual selection was further required, so we opted to use the MESH relations which are intensively used on medical articles classification.

The MT-based CLIR performance is closely related to the translation quality. Measuring MT quality frequently implies the use of human resources through objective and subjective metrics which are costly and time consuming. So we take use of the algorithms NIST/BLEU and the METEOR, for evaluating our dictionary development by scoring those machine translated texts it produces, during its developmental phase. These algorithms do not consist of measures of MT quality. Quality measures involve the use of human's metrics while NIST/BLEU algorithms measure the similarity between documents, and METEOR calculates precision and recall, as well as derived measures. In relation to NIST/BLEU algorithms, low scores do not necessarily result from low quality translations. Nonetheless, high scores are more indicative of good quality translations. These scores are absolute and widely vary between the diverse studies (NIST, 2001; Papieni, 2002; Culy, 2003). Although METEOR scores lie between zero and one, we still use the HIGH and LOW groups for comparative purposes. BLEU/NIST scores do not have an upper limit, since the values are proportional to the number of references used.

Our study is carried through a specialized multilingual parallel dictionary presenting low implementation cost and can be easily reproduced. Morphological distinct strings, in diverse languages and from different sources, are grouped under one single concept, making possible its use as translation equivalents. This permits the comparison between our MT groups, the DRY and the TEST ones, each of them representing two different stages, before and after the dictionary incorporation. They are escorted by other two scores HIGH and LOW, representing the theoretical upper and lower limits of machine translation performance. The HIGH group is composed of English strings which are morphologically and semantically correct. The LOW group consists of strings semantically equivalent but morphologically different, since they are not in English. Despite the intrinsic incorrectness of this group, we obtained scores above nullity, probably an effect of the similarity between words from diverse languages on the medical domain. This illustrates the main characteristic of this experiment: the capability to use a multilingual controlled resource to generate string sets for MT and string sets as references for evaluation purposes. We use the multilingual structure of UMLS to assist MT on medical applications.

The performance of machine translated texts without the use of the specialized dictionary

(DRY group) was greater than LOW group scores, representing the minimum performance of the MT system. In a previous study, we manually evaluated MT on clinical texts of thorax x-rays with the same system configuration without the use of a dictionary. We scored the ratio of correct translated words of the total words, and the ratio of intelligible sentences without word errors. There were 89% of correctly translated words and 67% of intelligible sentences (Castilla, 2004).

The performance of group TEST (machine translated with dictionary) was greater than the DRY as expected. There were significant improvements of the MT scoring after the incorporation of the specialized dictionary, demonstrated in all tested groups by both evaluation methodologies. The METEOR F-mean score was over 50% on all evaluated groups indicating favorable results.

Since we had favorable results on MT of thoracic radiology reports without using our specialized dictionary, we can expect with these results a significant performance improvement after its incorporation.

5 Conclusions and Future Work

The UMLS was successfully used as a substrate for testing and as a reference parameter for automated MT evaluation on medical terms. We also could demonstrate significant performance improvement after incorporating our specialized dictionary into a commercial MT system. Our proposed methodology can be reproduced to evaluate other MT systems or algorithms. The concept selection can be easily fitted to other domains of medical terminology. The concluding step of this project's phase is human evaluation of MT of thoracic radiology reports using our specialized dictionary.

Finally, we believe we increased the knowledge of MT on medical texts. The automated translation of medical texts and even medical speech could certainly be a significant tool in international humanitarian aid.

References

A.C. Castilla, S.S. Furuie. 2004. Avaliação da Tradução Automatizada de Relatórios de Radiografias de Tórax: Resultados Preliminares. Poster presented on the 9th Congress of Sociedade Brasileira de Informatica em Saude, Ribeirao Preto, Brazil

M.Y. Chang, Y.C. Sun, C.F. Chang, et al. 2003. A Free Radiology Dictionary Made From Abstract Corpus Of The Radiology And Radiographics. http://www.vghtpe.gov.tw/~rad/rsna2003/

C. Culy, S. Z. Riehemann. 2003. The Limits of N-gram Translation Evaluation Metrics. MT Summit IX, New Orleans, USA.

M. Dillinger. 2001. Dictionary development workflow for MT: design and management. MT Summit VIII, Santiago de Compostela, Spain, pp.83-87.

M. Eck, S. Vogel, A. Waibel. 2004. Improving Statistical Machine Translation in the Medical Domain using the Unified Medical Language system. In Proceedings of Coling 2004, Geneva, Switzerland.

E. Hovy, M. King, A. Popescu-Belis. 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 16:1-33.

A. Lavie , S. Banerjee. 2005. The METEOR Automatic Machine Translation Evaluation System. http://www-2.cs.cmu.edu/~alavie/METEOR/

A. Lavie, K. Sagae, S. Jayaraman. 2004. The Significance of Recall in Automatic Metrics for MT Evaluation. Preceedings of the 6th Conference of the Association for Machine Translation in the Americas, Washington, DC.

National Institute of Standards and Technology. 2001. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

D. W. Oard, B.J. Dorr. 1996. A Survey of Multilingual Text Retrieval. *Computer Science Technical Report Series.* Vol. CS-TR-3615.

K. Papineni, S. Roukos, T. Ward, W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics, pp. 311-318.

Radiological Society of North America. 2005. RADLEX: A Lexicon for Uniform Indexing and Retrieval of Radiology Information Resources. http://www.rsna.org/radlex/

S. Vintar, P. Buitelaar, M. Volk. 2003. Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval. International Workshop on Adaptive Text Extraction and Mining , Dubrovnik – Croatia.