IMPROVING A JAPANESE-SPANISH MACHINE TRANSLATION SYSTEM USING WIKIPEDIA MEDICAL ARTICLES

Jessica C. Ramírez[1,2], Yuji Matsumoto[2] and Darwin Muñoz[1]

[1]Universidad Iberoamericana, UNIBE, Santo Domingo, Dominican Republic
j.ramirez1@unibe.edu.do, d.munoz@unibe.edu.do
[2]Information Science, Nara Institute of Science and Technology, Nara, Japan
matsu@is.naist.jp

ABSTRACT

The quality, length and coverage of a parallel corpus are fundamental features in the performance of a Statistical Machine Translation System (SMT). For some pair of languages there is a considerable lack of resources suitable for Natural Language Processing tasks. This paper introduces a technique for extracting medical information from the Wikipedia page. Using a medical ontological dictionary and then we evaluate on a Japanese-Spanish SMT system. The study shows an increment in the BLEU score.

KEYWORDS

Comparable Corpora, Dictionary, Ontology, Machine Translation

1. INTRODUCTION

The quality, length and coverage of a parallel corpus are fundamental features in the performance of any Statistical Machine Translation (SMT) System. For some pair of languages there are a lack of aligned resources suitable for Natural Language Processing (NLP) tasks.

The use of automatic and semi-automatic methods for constructing resources along with manual resources help to reduce both the cost and time of any NLP project. For this reason many approaches have been published for constructing resources such as dictionaries, thesauri and ontologies, in order to facilitate NLP tasks such as word sense disambiguation, machine translation and other tasks [4]. [1] explore the multilingual features of Wikipedia for automatically extract sentences across multiple languages and [2] use Wikipedia for extracting Name Entities.

This study we use Wikipedia for extracting medical information from the health related articles in Japanese and Spanish, to construct a Medical Ontological dictionary, aligned the sentences in those articles and then we evaluate it impact on a Japanese-Spanish SMT system.

David C. Wyld et al. (Eds) : CCSIT, SIPP, AISC, NLP - 2015
pp. 111–116, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50411

2. RESOURCES

Wikipedia is an online multilingual encyclopedia with articles on a wide range of topics, in which the texts are aligned across different languages. Wikipedia have articles aligned in Spanish and Japanese. Wikipedia has some features that make it suitable for research such as:

Each article has a title, with a unique ID. “Redirect pages” handle synonyms, and “disambiguation pages” are used when a word has several senses. “Category pages” contain a list of words that share the same semantic category. For example the category page for “Birds” contains links to articles like “parrot”, “penguin”, etc. Categories are assigned manually by users and therefore not all pages have a category label.

The information in redirect pages, disambiguation pages and Category pages combines to form a kind of Wikipedia taxonomy, where entries are identified by semantic category and word sense.

3. GENERAL DESCRIPTION

The general goal is to extract useful in domain data from Wikipedia to improve the performance of a Japanese-Spanish SMT system. The study is divided 3 phases : The first one the construction of the Japanese-Spanish ontological dictionary, then Japanese-Spanish parallel and then evaluate the corpus in a SMT system.

Phase 1. Ontology Medical Dictionary

The goal is the creation of a Spanish-Japanese ontology, in which, we align each medical article in Spanish and Japanese and then, we extract all the terms related to the article title. And then by using Pattern Recogniton tecniques.We extract information associated to the given word, for example : Kidney Stone, desease related to kidney, symptons, causes, etc.

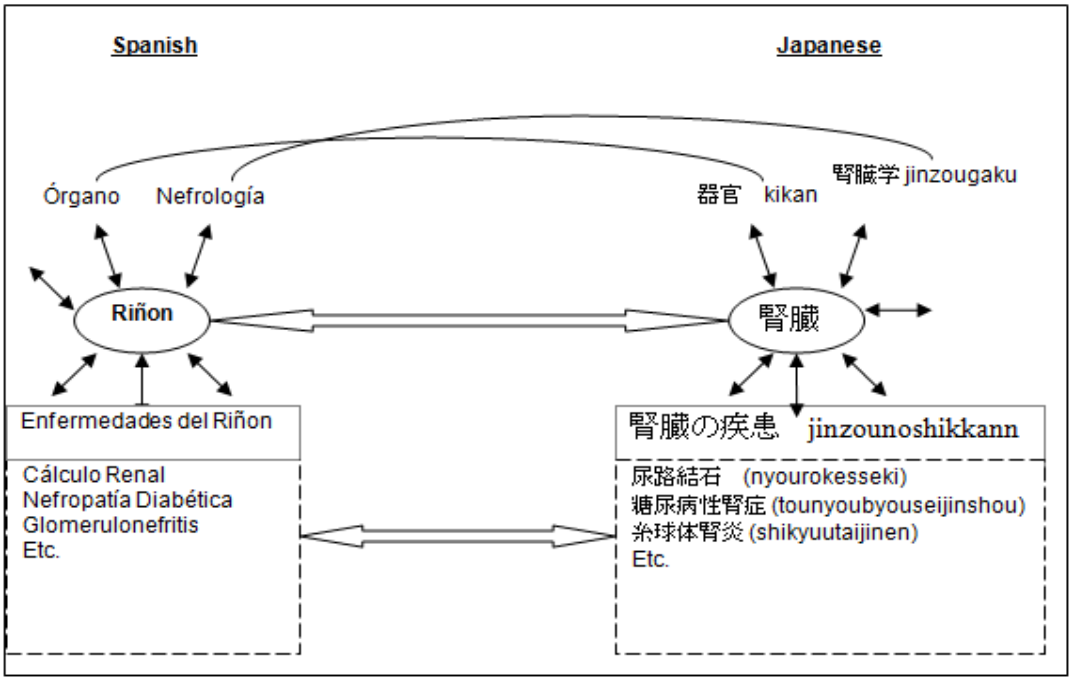

Figure 1 : General Structure of the Spanish-Japanese alignment.

Ex. In figure 1 [6] shows the structure of the system, for example, the word 'Kidney', Spanish is "Riñon", is translated to japanese a 腎臓, 'yinzo', which is associated with all the deseases related to kidney such as 'Kidney desease and a list of the desease such as :'Kidney stone', 'Glomerulonephritis', etc.

Methodology

Extracting The Medical articles from Wikipedia

The goal is acquisition of Spanish-Japanese medical domain of Wikipedia's article titles. Each Wikipedia article provides links to corresponding articles in different languages.

Every article page in Wikipedia has on the left hand side some boxes labelled: 'navigation', 'search', 'toolbox' and finally 'in other languages'. This has a list of all the languages available for that article, although the articles in different languages do not all have exactly the same contents.

To extract the medical articles we extract them by mean of their categories, mining all articles that belong to categories such as: "medicine", "disease", "organ", etc.

Pre-processing Procedure

We eliminated all irrelevant information of each article in Spanish such as tables, special characters, menus, etc.

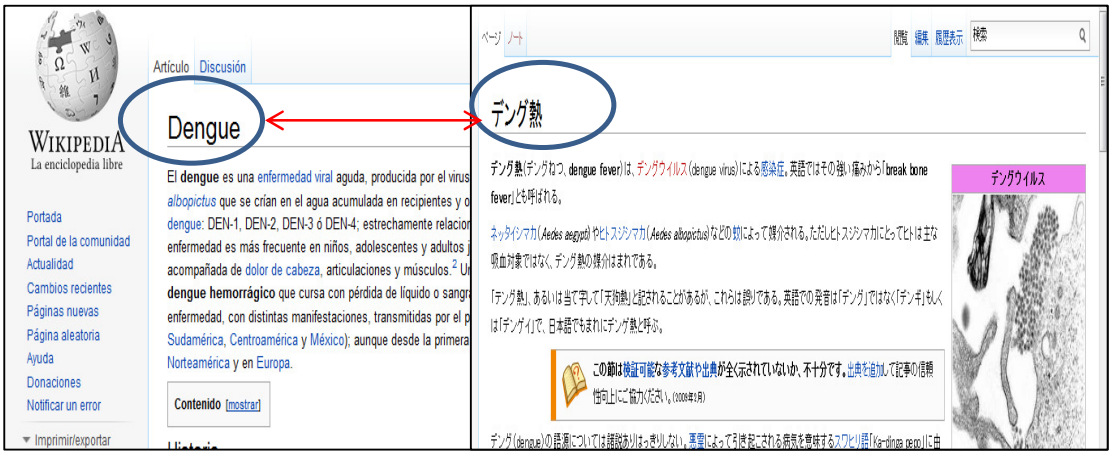

Figure 2 :Wikipedia links.

Dictionary Spanish-Japanese

Take all article titles that are nouns or named entities and look in the articles' contents for the box called '*In other languages*'. Verify that it has at least one link. If the box exists, it links to the same article in other languages. Extract the titles in these other languages and align them with the original article title.

For instance Figure 2 shows the Spanish article titled "Dengue" (Dengue fever), which is translated into Japanese as "デング熱"(DenguNetsu). When we click Spanish or Japanese '*in*

other languages' box we obtain an article about the same topic in the other language. This gives us the translation and we proceed to extract it.

Ontological Relation between terms

The goal is extract all features related to a given disease such as: symptoms, causes, organs, etc. By using Pattern Recognition we extract the sentences associated with the inner titles in Wikipedia. Ex. The inner title symptom in Spanish and Japanese, and extract the phrases and translated the nouns if there are hyperlinks and belong to the extracted Wikipedia dictionary. And we proceed to aligned them in both languages.

The long term goal pursue when we type a symptom like "headache", display all the diseases that that contain headaches in the list of symptoms, in case there add another symptoms like "fatigue" continue pruning the list with possible disease.

Phase 2. Constructing a Parallel Corpus

The goal is the creation of a parallel corpus by aligning the sentences of the medical articles. We use a extended form of a ruled-based approach similar to [5]. We extended the amount of rules and eliminate some rules that were redundant or cause ambiguity between rules.

Methodology

We eliminate the irrelevant information from Wikipedia articles, to make processing easy and faster.

The steps are as follows.

1. Remove from the pages all irrelevant information, such as images, menus, characters such as: "()", "&quot", "*", etc...
2. Verify if a link is a redirected article and extract the original article
3. Remove all stopwords -general words that do not give information about a specific topic such as "the", "between", "on", et

For splitting the sentences in the Spanish articles we used NLTK toolkit[1] , which is a well-known platform for building Python scripts.

For tag Spanish sentences, we used FreeLing[2], which an open source suit for language analizer, specialized in Spanish language.

For Splitting into sentences, in to words and add a word category, we used MeCab[3], which is a Part-of-Speech and Morphological Analyser for Japanese.

	Rule Description
Rules	**Japanese=> Spanish**
Noun	Noun+desu => noun

[1] http://nltk.org/
[2] http://nlp.lsi.upc.edu/freeling/
[3] http://cl.naist.jp/~eric-n/ubuntu-nlp/dists/hardy/japanese/

Name Entity	NE=>NE (Capital letter)
Adjective	Adj (fe/male) =>Adj (NA/I)
Question	(sentence+か)=>(¿ + sentence +?)
Pronouns	Pron =>Pron

Table 1. shows some of the rules applied to this work. Those rules are created taking in account the morphological and syntactic characteristic of each language.

Phase 3. Using Medical Corpus in to SMT System

The main goal is to measure feasibility of using an in-domain corpus (in this case health related) in a SMT system.

We used the aligned parallel sentences extracted in phase 2 to measure its impact in a Statistical Japanese-Spanish MT system.

Experiments

We use a random sample of 500 parallel sentences extracted from Wikipedia and we add to 50k Japanese-Spanish parallel corpus from Europarl. We used human translators to translate into Japanese the 50k sentences because the Europarl corpus just contains the proceedings of the European Parliament for countries that belongs to the European Union.

We train a baseline SMT system with the 50k sentences. Then we performed experiments adding the 500 sentences extracted on phase 2 to the baseline. In both cases we used as a language model Wikipedia Spanish articles, 10k for development set, 10k for test set and 30k for training.

4. RESULTS AND DISCUSSION

Table 2 shows the results using the Europarl data and the result by adding the the medical corpus. Using the Medical corpus increase the BLEU[4] score. However, If in the training set there is not health related sentences the BLEU score do not increase.

Corpus	BLEU
EuroParl	27.87%
EuroPal + Medical corpus	28.15%

Table 2. Results

5. CONCLUSIONS

This paper focuses on extracting medical information from Wikipedia and the creation of an ontology in Spanish and Japanese. In domain corpus can be used to improve the performance of a SMT system.

We will extend this work by using several corpus of other field, like economy, sociology and so on.

ACKNOWLEDGEMENTS

[4] Bilingual Evaluation Understudy

This research was supported by is « Fondo Nacional de Innovación y Desarrollo Científico y Tecnológico » **FONDOCyT#2012-2013-3A2-59,** Santo Domingo, Dominican Republic.

We would like to thanks to Yuya R.

REFERENCES

[1] Adafre, Sisay F. & De Rijke, Maarten, (2006) "Finding Similar Sentences across Multiple Languages in Wikipedia", In *Proceeding of* EACL-06, pages 62-69.

[2] Bunescu, Razvan & Pasca, Marius (2006) "Using Encyclopedic Knowledge for Named Entity Disambiguation", In *Proceeding of* EACL-06, pages 9-16.

[3] Fung, Pascale & Cheung Percy, (2004) " Multi-level Bootstrapping for extracting Parallel Sentences from a quasi-Comparable Corpus", In *Proceeding of the 20th International Conference on Computational Linguistics*. Pages 350

[4] Ramírez, Jessica, Asahara, Masayuki & Matsumoto, Yuji , (2008) "Japanese-Spanish Thesaurus Construction Using English as a Pivot", In *Proceeding of The Third International Joint Conference on Natural Language Processing* (IJCNLP), Hyderabad, India. pages 473-480.

[5] Ramírez, Jessica & Matsumoto, Yuji (2012) "A Ruled-Based Approach for Aligning Japanese-Spanish Sentences from a Comparable Corpora", *International Journal of Natural Language Computing* (IJNLC).

[6] Ramírez, Jessica & Matsumoto, Yuji (2013) "Extracción Automática de Diccionario Médico Japonés-Español. *Actualizaciones en Comunicación Social*, Santiago, Cuba. Vol.II.

[7] Ramírez, Jessica,Matsumoto, Yuji, Muñoz, Darwin & Joyanes, Luís (2013) "Construcción Automática de Corpus Paralelo Japonés-Español en el Área de la Salud. *Memorias de VIII Conferencia Internacional de Lingüística*, Habana, Cuba.

AUTHORS

Jessica C. Ramírez

She received his M.S. degree from Nara Institute of Science and Technology (NAIST) in 2007. She is currently pursuing a Ph.D. degree. Her research interest Include machine translation and word sense disambiguation.

Yuji Matsumoto

He received his M.S. and Ph.D. degrees in information science from Kyoto University in 1979 and in 1989. He is currently a Professor at the Graduate School of Information Science, Nara Institute of Science and Technology. His main research interests are natural language understanding and machine learning.

Darwin Muñoz

He received his M.S. degree in Business from Quebec University, Canada in 2004 and Ph.D. degree in information science from Universidad Pontificia de Salamanca, Spain in 2014. He is currently a Professor at Universidad Iberoamericana.