Procesamiento del Lenguaje Natural, Revista nº 52 marzo de 2014, pp 29-36 recibido 14-11-13 revisado 04-02-14 aceptado 05-02-14

Collecting and POS-tagging a lexical resource of Japanese biomedical terms from a corpus

Recogida y etiquetado morfológico de un lexicón de términos biomédicos en japonés a partir de corpus

Carlos Herrero Zorita, Leonardo Campillos Llanos, Antonio Moreno Sandoval
Laboratorio de Lingüística Informática. Departamento de Lingüística
Facultad de Filosofía y Letras. Universidad Autónoma de Madrid
c\ Francisco Tomás y Valiente, 1. Campus de Cantoblanco. Madrid 28049
{carlos.herrero, leonardo.campillos, antonio.msandoval}@uam.es

Resumen: El artículo resume el proceso de recopilación de un lexicón de términos biomédicos en japonés etiquetados morfológicamente. En primer lugar se han considerado para esta tarea las características morfosintácticas del japonés así como el origen y formación de los términos médicos en esta lengua. Posteriormente la lista se ha recopilado utilizando el corpus japonés MultiMedica, las etiquetas especiales de un etiquetador morfológico y varios diccionarios médicos especializados. Para el siguiente proceso de etiquetado se han considerado tres etiquetadores japoneses (ChaSen, Mecab, Juman), de los cuales se ha escogido este último. Una vez etiquetado, se ha corregido el problema de la *sobresegmentación* de los términos japoneses y se han simplificado las etiquetas para el propósito de nuestra tarea. Este recurso es la base para la creación de un extractor de términos médicos en japonés.

Palabras clave: terminología médica, japonés, recurso léxico, análisis morfológico.

Abstract: The following paper explains the methodology followed for the creation of a morphologically tagged medical lexicon in Japanese. In order to build this medical resource we have taken into account the morphosyntactic characteristics of the language as well as the origins and formation of the medical terms. Following this, we have compiled a list using the Japanese MutiMedica corpus, special tags from a POS tagger, and several specialised medical dictionaries. After considering three different taggers (ChaSen, Mecab, Juman) we finally chose Juman for the tagging of the lexicon. The problem of the *oversegmentation* of the language was then corrected and the tags have been normalised. This resource is the base component for the creation of a medical term extractor.

Keywords: medical terminology, Japanese, lexical resource, POS tagging.

1 *Introduction*

Natural language processing tasks for domain-specific texts (e.g. biomedicine) rely upon comprehensive lexical resources. The Unified Medical Language System (UMLS) Specialist lexicon and Metathesaurus (Donnelly, 2006) are the major resources available for English (Bodenreider, 2006 provides further references). Despite the lack of thesaurus for other languages, multilingual lexical databases such as EuroWordNet have also been applied in the field of medical terminology (Vivaldi and Rodríguez, 2002).

In the following pages we will present a morphologically tagged Japanese lexicon of the medical domain. This list of terms will be used for developing a Japanese automatic term recognition (ATR) system.

The article explains the methodology followed towards the creation of the lexicon. For this purpose, we took two steps: firstly, we translated 548 Graeco-Latin medical affixes into Japanese; and secondly, we compiled the medical lexicon. The list of terms was compiled using three resources: a medical corpus, Japanese morphological analysers, and specialised dictionaries. The lexicon was morphologically tagged, taking into account the morphosyntactic challenges that this language entails.

The paper is divided as follows. Section 2 will describe the MultiMedica project and corpus. Section 3 will provide a theoretical background, including the formation of Japanese medical terms and the origins of the medical terms in Japanese and in Western languages. Finally, Section 4 will explain the steps given towards the creation of the lexicon.

2 Description of the MultiMedica project and corpus

The data of this work is based on the Japanese texts from the MultiMedica corpus. This collection was compiled by the Computational Linguistics Laboratory at the Autonomous University of Madrid (LLI-UAM)[1], as part of the MultiMedica project (Martínez et al., 2011)[2]. It is a specialised comparable corpus formed by biomedical texts written in Spanish, Arabic, and Japanese.

The corpus assembles 51,476 documents and more than seven and a half million words in three languages, Arabic, Japanese and Spanish (Moreno-Sandoval and Campillos-Llanos, 2013). Documents were gathered from professional books and journals that were written by medical doctors, as well as from articles drafted by health professionals and edited by journalists. Thus, the corpus collects both technical and informative articles. Texts cover most medical specialties.

2.1. The Japanese corpus

The Japanese corpus is made up of abstracts from medical journals on different specialties (e.g. Oriental Medicine, Obstetrics, and Gynecology). The total corpus size is 1,131,304 Japanese characters (kanji and kana) (Table 1).

2.2. The query interface

The following step of the project at the LLI-UAM was the development of a query interface that allows the user to consult and concordance the corpus. This tool, which is still in beta phase, will allow multiple search options, including the distinction between form and lemma, beginnings and endings of words, and morphological category, as well as other functionalities such as frequency extraction and collocations. Figure 1 shows an example of the result of a query made of the kanji 肝 ('liver') in the Japanese corpus. The next section will provide a brief description of previous work related to this matter, as well as the origins of the medical terms in Spanish, English, and Japanese.

Japanese corpus	Texts	Characters
Kampo Medicine (Oriental medicine in Japan)	719	214,757
Kansenshogaku Zasshi (Infectious diseases Journal)	858	244,879
Kanzo (Liver diseases Journal)	1,446	432,674
ORLTokyo (Japanese otolaryngology)	623	203,705
Sanfujinka no shinpo (Advances in obstetrics)	100	35,289

Table 1: Description of the Japanese corpus

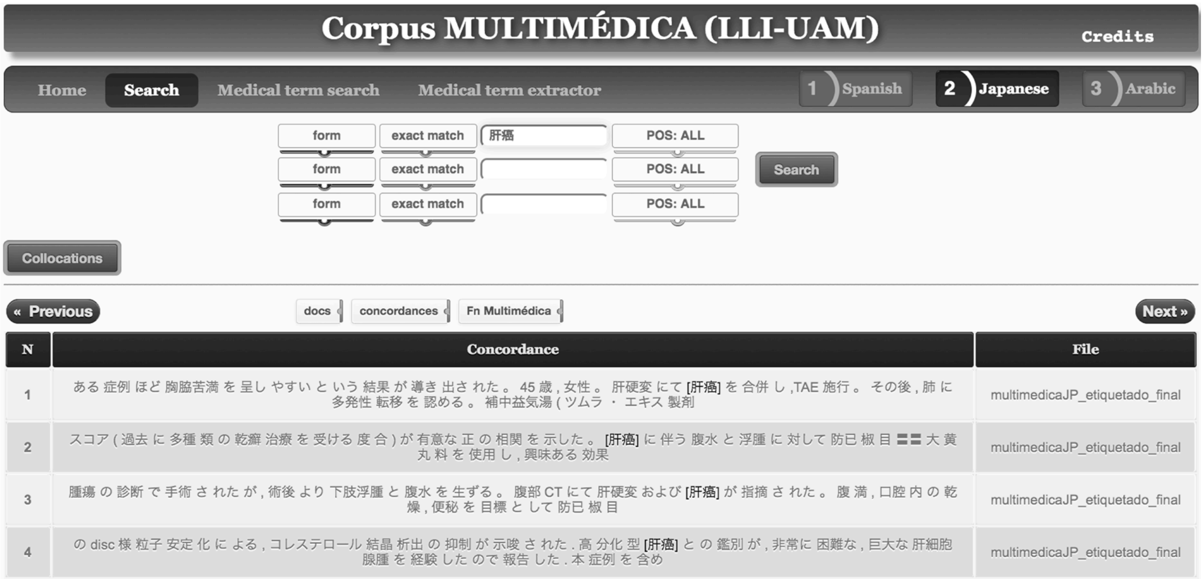
3. Theoretical background

3.1. Previous work

Japanese proves to be a challenging language for Natural Language Processing tasks, due to its morphosyntactic characteristics. One of the most noticeable problems is the so-called *oversegmentation* (Hisamatsu and Nitta, 1996). Morphological taggers have problems to tokenize Japanese words due to the fact that characters are not separated by blank spaces, and the agglutinative nature of this language. This is especially problematic in formal or specialised discourses.

[1] http://www.lllf.uam.es/ING/Multimed.html

[2] MultiMedica is a coordinated project involving the Universidad Carlos III de Madrid, the Universidad Politécnica de Madrid, and the Universidad Autónoma de Madrid. It aims at research on natural language processing for the biomedical domain. Further information is available at: http://labda.inf.uc3m.es/multimedica/.



form | exact match | 肝癌 | POS: ALL
form | exact match | | POS: ALL
form | exact match | | POS: ALL

N	Concordance	File
1	ある 症例 ほど 胸脇苦満 を 呈し やすい と いう 結果 が 導き 出さ れた 。 45 歳 , 女性 。 肝硬変 にて [肝癌] を 合併 し ,TAE 施行 。 その後 , 肺 に 多発性 転移 を 認める 。 補中益気湯 (ツムラ ・ エキス 製剤	multimedicaJP_etiquetado_final
2	スコア (過去 に 多種 類 の 乾癬 治療 を 受ける 度 合) が 有意な 正 の 相関 を 示した 。 [肝癌] に 伴う 腹水 と 浮腫 に 対して 防已 椒 目 〓〓 大 黄 丸 料 を 使用 し , 興味ある 効果	multimedicaJP_etiquetado_final
3	腫瘍 の 診断 で 手術 さ れた が , 術後 より 下肢浮腫 と 腹水 を 生ずる 。 腹部 CT にて 肝硬変 および [肝癌] が 指摘 さ れた 。 腹 満 , 口腔 内 の 乾燥 , 便秘 を 目標 と して 防已 椒 目	multimedicaJP_etiquetado_final
4	の disc 様 粒子 安定 化 に よる , コレステロール 結晶 析出 の 抑制 が 示唆 さ れた . 高 分化 型 [肝癌] と の 鑑別 が , 非常に 困難な , 巨大な 肝細胞 腺腫 を 経験 した ので 報告 した . 本 症例 を 含め	multimedicaJP_etiquetado_final

Figure 1: The search interface

This problem has been widely studied in works concerning compound word analysis and recognition of unknown terms (Nagata 1999; Masaaki,1999; Han et al. 2002; Kudo, 2007; Murawaki and Kurohashi, 2010, among others) with different approaches towards the development of automatic term extractors (Nakagawa and Mori, 2002 and Oh et al. 2000).

As Murawaki and Kurohashi (2010: 832) explain, dictionaries are indispensable for Japanese morphological analysis because not only part-of-speech (POS) tagging is required, but also a process of segmentation. In our project, we have collected a Japanese medical lexicon that has been morphologically tagged. Following this, the oversegmentation has been manually corrected. The next section provides a reflection on Japanese medical terminology and on the challenges that arose when we processed it in comparison to other languages.

3.2. Formation and types of terms

Medical terms in Western languages have their origins in Ancient Greece in the Hippocratic Corpus. These terms would then be adapted in Rome by Galen, whose medical practice would dominate the medical knowledge until the beginning of the modern era (Longrigg, 2002: 29-39). For this reason, even though medical practices have changed today, the language—or, more specifically, the language of medical technicalities—still has its origin in ancient Greek and Latin. Terms are, therefore, formed by the addition of Graeco-Latin affixes. For example, *gastritis* (‘inflammation of the lining of the stomach’) is constructed with the root *gastr-* (from Latin *gastro-*, ‘stomach’, which originally evolved from Greek *grastro-*), and with the Graeco-Latin suffix *-itis* (‘diseases characterized by inflammation’).

In Japanese, on the other hand, the picture is quite different. From the early beginnings of the Japanese culture, Chinese medicine was the major way of medical practice in Japan. Hence, terms belong to Chinese characters, adapted over the years into the Japanese kanji (Izumi and Isozumi, 2001: 91). However, the vast majority of the medical terms employed today were borrowed from Western languages. Since the Sakoku era, the first medical terms from the West arrived through the medicine books traded with Dutch merchants (Irwin, 2011: 37). These words finally rooted officially in the language in the 19th century, when Japan opened up and the Meiji government adopted the German medical educational system (2011: 51). The initial loanwords were introduced by means of two different processes: on the one hand, (1) the translation and coining into Sino-Japanese compounds using kanji and, on the other, (2) their transcription into the katakana[3] alphabet.

We will find, therefore, the following types of terms in our corpus:

[3] Japanese combines three writing systems: Kanji (ideograms of Chinese origin), hiragana (a syllabic system from Japan) and katakana (also a syllabic alphabet, mainly used for transcribing foreign words). A fourth, non-Japanese alphabet is used, named romaji, that uses Latin characters.

- Usage of Japanese kanji characters for Chinese Medicine and Western Medicine terms, e.g.[4]:

  卵管癌

  egg-tube-cancer

  ran-kan-gan

  'fallopian tube cancer'

- Transcriptions into katakana, e.g.:

  アキネジア

  akinesia

  a-ki-ne-ji-a

  'akinesia'

- Terms using both kanji and katakana, e.g.:

  遅発性ジスキネジア

  behind schedule-ness/dyskinesia

  chihatsu-sei/ji-su-ki-ne-ji-a

  'tardive dyskinesia'

- Borrowings: e.g. DNA

Since Japanese is an agglutinative language, we can assume that the majority of terms written in kanji will be formed by composition using free morphemes. This process is very different from affixation in English and Spanish.

4. Methodology and results

4.1. Resources

For the development of this project, we used several tools for the processing of the Japanese language. First of all, we used the Juman[5] morphological analyzer, developed at the Kurohashi Lab of Kyoto University. We also considered using the ChaSen[6] and Mecab[7] taggers. As we will see in Section 4.4, we selected Juman for this purpose due to the extensive morphological information it provides from each word, such as specialised tags that automatically recognises medical and anatomy terms. Secondly, we used two medical dictionaries, the *Online Life Science Dictionary*, which belongs to the Life Science Project developed at Kyoto University, and the *Japanese-English-Chinese Dictionary*, from 朝倉書店 publications (1994).

Our work can be divided into two steps: (1) the translation and analysis of medical Graeco-Latin affixes from English to Japanese, and (2) the creation of a dictionary of Japanese medical terms. Both stages will be conditioned by the agglutinative nature of Japanese language, and the fact that there are no blank spaces between words.

4.2. Translation of medical affixes

Using medical affixes for recognising medical terms has ended in a high level of precision (Estopà et al. 2000, Moreno-Sandoval et al., 2013). We took this approach for Japanese: the starting point was a list of 467 Graeco-Latin medical affixes collected by the LLI-UAM. Each of them was translated into Japanese using the online Japanese-English medical dictionary *Online Life Science Dictionary*, which allows the user to search for beginnings and endings of words. We discarded those that did not appear in the corpus. Afterwards, we tagged them using the Juman morphological analyser.

We observed that this step would not achieve the same results as in Spanish. First of all, the Japanese medical terms are predominantly formed by composition, adding free morphemes[8], instead of affixation (Herrero-Zorita, 2013) (See Figure 2). Only a 7.30% are affixes (3.47% prefixes, and 3.83%, suffixes). These include affixes that do not necessarily belong exclusively to the medical domain. Secondly, these free morphemes do not classify the word into a medical term. Whereas a word containing *cardio-* in Spanish will refer to a term related to the heart, the translation of this prefix into Japanese results in the free morpheme 心 'heart, mind'. Although the word is equally used for medical terms, e.g. 心疾患 ('cardiopathy'), it is also used in order compounds that do not necessary belong to a medical term, for example: 心酔 ('adoration').

[4] The following examples include: (1) the word in Japanese, (2) the literal translation, (3) the reading in romaji and (4) the translation in English. In the case of katakana words, (2) and (4) overlap, since they are phonological transcriptions.

[5] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[6] http://chasen-legacy.sourceforge.jp/

[7] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[8] Free morphemes, morphemes that can stand alone as independent words, are differentiated from prefixes and suffixes (bound morphemes) that appear as part of a larger word.

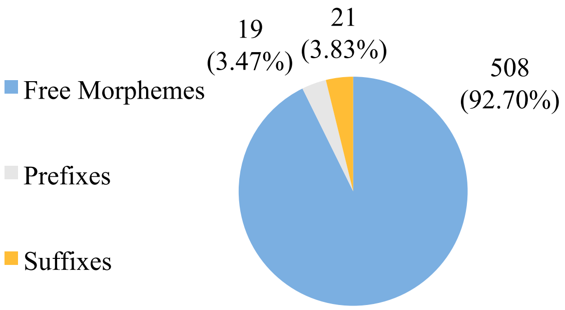


Figure 2: Japanese morphemes using translated Graeco-Latin affixes according to Juman

4.3. Compilation of the lexicon

The solution was to create a lexicon of Japanese medical terms and assign to each of them a grammatical category. We compiled by hand a list of 31,458 terms taken from the two specialised dictionaries previously mentioned. Following this, we completed this listing with words that were extracted from the corpus, automatically recognised by means of the specialised tags included in Juman. Figure 3 represents an example of the list (the translation has been provided in this paper for the sake of clarity). Table 2 shows the distribution of the terms according to the writing system. We can see that the usage of kanji is predominant.

…
前糖尿病状態 ('prediabetic state')
精神病状態 ('psychotic state')
前腫瘍状態 ('preneoplastic condition')
前癌状態 ('precancerous condition')
…

Figure 3: Sample of the Japanese lexicon

Writing System	Terms	%
Kanji	23,373	74.299
Kanji + Katakana	6,028	19.162
Katakana	2,002	6.364
Borrowings	40	0.127
Hiragana	15	0.004

Table 2: Distribution of the terms

4.4. Tagging process

The list was then tagged, since we needed additional linguistic information from the terms.

There are three widespread taggers in Japanese that we considered using for this task: the Juman, ChaSen, and Mecab. The main problem in this step was *oversegmentation*. In this case, medical terms formed by two or more kanji that do not appear in common dictionaries are split into recognisable morphemes. The degree of the segmentation depends on the tagger. For example, the term 肝生検 'liver biopsy' is divided as the following:

- Juman splits it into two terms: 肝 ('liver') and 生検 ('biopsy').
- ChaSen and Mecab, into three words: 肝 ('liver'), 生 ('raw'), and 検 ('examination').

Also, each tagger provides different degrees of linguistic information. To choose the appropriate tagger, we carried out three comparisons between the three taggers. We took into account the problem of oversegmentation and the morphological information provided.

First, we tagged the MultiMedica corpus using each program. A word list was obtained, and we looked up the terms in our lexicon (Table 3). Secondly, we tagged the lexicon and observed how many words were generated by each tagger after the segmentation (Figure 4). Thirdly, we observed the degree of information given by each one of them. We take as an example the word 学校 ('school') (Table 4).

	Types in word list	Words found in lexicon	%
ChaSen	10,020	2,358	23.53
Juman	10,819	2,334	21.57
Mecab	11,575	2,484	21.46

Table 3: Terms from the corpus in the lexicon

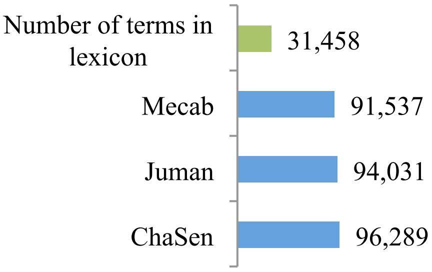


Figure 4: Words obtained after tagging the lexicon

<table>
<tr><th>ChaSen</th></tr>
<tr><td>学校 | ガッコウ | 学校 | 名詞 | 一般
Word | Reading | Lemma | Tag | Subtag</td></tr>
<tr><td><b>Mecab</b></td></tr>
<tr><td>学校 | 名詞 | 一般 | *,*,*,* | 学校 | ｶﾞｯｺｳ | ｶﾞｯｺｰ
Word | Tag | Subtag | Lema | Reading | Reading variation</td></tr>
<tr><td><b>Juman</b></td></tr>
<tr><td>学校 がっこう | 学校 | 名詞 | 6 | 普通名詞 | 1 * 0 * 0 | "代表表記:学校/がっこう | カテゴリ:場所-施設 | ドメイン:教育・学習"
Word | Reading | Lemma | Tag | Subtag | Reading variation | Domain</td></tr>
</table>

Table 4: Information given by each tagger

The three taggers retrieve similar results regarding the segmentation of the words. However, Juman provides a wider range of morphological information, including the specialised tags indicating the domain of the words. For this reason, we have chosen it for the tagging of the lexicon.

After the tagging and segmentation were completed by Juman, we re-*joined* the morphemes creating once again the complete term. That means 63,079 morphemes were erroneously split and corrected, a 66.72% of the total (94,031) (see Section 4.5). Then, we assigned the category given to the morpheme at the further right, and finally we translated the category. Figure 5 shows an example with the word 下垂体 ('pituitary gland'):

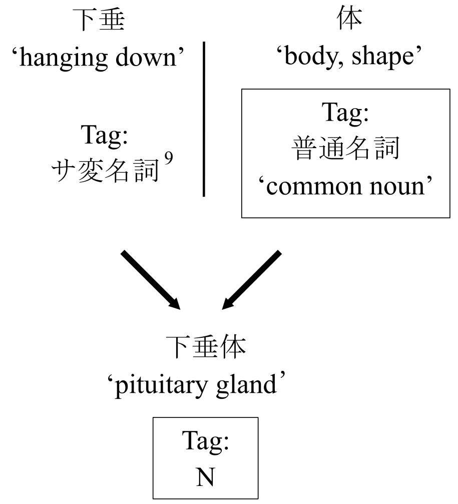


Figure 5: Correction and tagging of the segmentation

Since Japanese is a right-headed language (Miyaoka and Tamaoka, 2005: 46), the head situated at the right position determines the category of the complete compound. Through this procedure, we created a dictionary of Japanese terms that were morphologically tagged. This list includes long terms such as 自己分泌型細胞運動刺激因子受容体 ('Auto-crine Motility Factor receptor')[10]. These types of terms would have been segmented and not recognised automatically by any of the three morphological analysers (Figure 6).

上強膜炎 ('Episcleritis')	N
等電点 ('Isoelectric point')	N
原虫類 ('Protozoa')	N
偽動脈瘤 ('Pseudoaneurysm')	N
手術不能 ('Inoperable')	N | ADJ
c-Metタンパク質 ('C-Met protein')	N
細胞容積 ('Cell volume')	N
組織化学的 ('Histochemical')	ADJ
遺伝子導入 ('Gene transfer')	N
細胞質体 ('Cytoplast')	N
坐薬('Suppository')	N
植え込む ('To implant')	V
ウイルス組み込み('Virus integration')	N
第12脳神経 ('Twelfth cranial nerve')	N

Figure 6: Sample of the Japanese lexicon (including translation)

4.5. Dealing with oversegmentation

Following this, the compilation of such lexicon allowed us to correct the medical terms that were oversegmentated after the tagging of the MultiMedica corpus. For this purpose, we first

[9] サ変名詞, *sahenmeishi*, is a type of noun that can be attached to the auxiliary *suru* ('to be') to form a verb. For example, from the noun *benkyou* ('stu-dy'), we can form the verb *benkyou-suru* ('to study').

[10] Formed by 自己分泌 (*jikobunbi*, 'autocrine') + 細胞運動 (*saibouundou*, 'cell mobility') + 刺激因子 (*shigekiinshin*, 'stimulating factor') + 受容体 (*juyoutai*, 'receptor').

looked for the terms from the lexicon that appear in the corpus (6,811 types, see Table 5). We then tagged the corpus with Juman. Lastly, we followed the same process as with the lexicon: we corrected the segmented terms and assigned them their POS tag. Table 5 shows the results of this operation:

	Types	% of the total corpus	% of the terms extracted
MultiMedica (tagged)	28,325	100.00	-
Lexicon terms in corpus	6,811	24.05	100.00
Corrections	5,739	20.37	84.36

Table 5: Results of correcting oversegmentation in the MultiMedica corpus

We can observe the overall importance of the oversegmentation problem: the tagger split more than 66% of the morphemes of the lexicon (Section 4.4); this led to a correction of around 84% of the terms extracted, a 20.37% of the total corpus. In other words, the reliability of the current taggers for Automatic Term Recognition is very low. Both outcomes should be taken into account when processing complex lexical units in non-segmenting languages such as Japanese or Chinese.

5. Conclusions and future work

In this paper we have presented a morphologically tagged lexicon of Japanese medical terms. First, we have explored the origins and formation of medical terms; secondly, we have presented the problems of the morphological segmentation; and finally, we have explained the process of compiling the lexicon.

From our experience, it seems imperative to take into account the morphosyntactic characteristics of Japanese when performing a natural language processing task—especially, when dealing with automatic tagging. The agglutinative nature of the language and the lack of white spaces between words are the main problems for these types of tasks. In order to compile the lexicon we have used two medical dictionaries and the special tags of the Juman tagger. After the tagging process, we have overcome the *oversegmentation* problem by manually joining together the separated morphemes, and have translated the tags to a universal codification.

This lexicon will not only serve as a lexical resource and a reliable source of information, as it will become the foundation for a medical automatic term extractor.

Acknowlegments

This research has been funded by the MINECO (under the grant TIN2010-20644-C03-03) and by the Madrid Regional Government (grant MA2VICMR).

Bibliography

Bodenreider, O. 2006. Lexical, Terminological and Ontological Resources for Biological Text Mining. In S. Ananiadou, and J. McNaught (eds.) *Text mining for biology and biomedicine,* 43-66. Boston: Artech House.

Donnelly, K. 2006. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. In L. Bos et al. (eds.) *Medical and Care Compunetics,* 3.

Estopà, R., J. Vivaldi, and Mª. T. Cabré 2000. Use of Greek and Latin forms for term detection. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000).* Athens, Greece.

Han, D., T. Ito, and T. Furugoori. 2002. A Deterministic Method for Structural Analysis of Compound Words in Japanese. *Language, Information and Computation: Proceedings of the 16th Pacific Asia Conference 2000.* Jeju, Korea.

Herrero-Zorita 2013. An initial approach on medical term formation in Japanese through the usage of corpora. *Proceedings of the 7th Corpus Linguistics Conference 2013*, 339-340, Lancaster University, Lancaster, United Kingdom), July.

Hisamitsu, T., and Y. Nitta. 1996. Analysis of Japanese compound nouns by direct text scanning. In *Proceedings 16th Conference on Computational Linguistics,* 1: 550-555. Stroudsburg, PA, USA.

Irwin, M. 2011. *Loanwords in Japanese.* Amsterdam: John Benjamins Publishing.

Izumi, Y., and K. Isozumi. 2001. Modern Japanese medical history and the European influence. *The Keio journal of medicine*, 50 (2): 91-99.

Japanese-English-Chinese dictionary. 1994. 朝倉書店 (Asakura Shoten)

Kudo, M. 2007. *A lexical semantic study of four-character Sino-Japanese compounds and its application to machine translation* PhD Thesis. Dept. of Linguistics - Simon Fraser University.

Longrigg, J. 2002. Medicine in the Classical World. In Loudon, I. (ed.) *Western Medicine.* Oxford: Oxford University Press.

Martínez, P., J.C. González Cristobal, and A. Moreno-Sandoval 2011. MULTIMEDICA: Extracción de información multilingüe en Sanidad y su aplicación a documentación divulgativa y científica. *Procesamiento del Lenguaje Natural,* 47, 347-348. Retrieved from http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/1003

Masaaki, N. 1999. A part of speech estimation method for Japanese unknown words using a statistical model of morphology and context. *Proceedings ACL*, 277–284.

Miyaoka, Y. and K. Tamaoka. 2005. Investigation of the Right-hand Head Rule Applied to Japanese Affixes. *Glottometrics,* 10: 45-54.

Moreno-Sandoval, A., Campillos-Llanos 2013. Design and Annotation of MultiMedica – A Multilingual Text Corpus of the Biomedical Domain. *Procedia - Social and Behavioral Sciences*, 95 (25): 33-39.

Moreno-Sandoval, A., L. Campillos-Llanos, A. González-Martínez, and J. M. Guirao 2013. An affix-based method for automatic term recognition from a medical corpus of Spanish. In *Proceedings of the 7th Corpus Linguistics Conference 2013*, 214-217, Lancaster University, Lancaster, United Kingdom), July.

Murawaki, Y. and S. Kurohashi. 2010. Online Japanese Unknown Morpheme Detection using Orthographic Variation. *LREC, European Language Resources Association*.

Nakagawa, H. and T. Mori. 2002. A simple but powerful automatic term extraction method. *Proceedings COMPUTERM 2002: second international workshop on computational terminology*, 14.

Oh, J-H., J. Lee, K-S. Lee, and K-S. Choi. 2000. Japanese term extraction using dictionary hierarchy and machine translation system. *Special Issue of Terminology*, 6 (2): 287-311.

Online Life Science Dictionary (ライフサイエンス辞書オンラインサービス). 2013. Available at: http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/index.html.

Vivaldi, J., and H. Rodríguez, 2002. "Medical Term Extraction using EWN ontology". In *Proceedings of Terminology and Knowledge Engineering 2002 (TKE'02)*. Nancy: 137-142.