Semantic transference for enriching multilingual biomedical knowledge resources

María Pérez[a,*], Rafael Berlanga[a]

[a]*Departament of Computer Languages and Systems, Universitat Jaume I, Avda. Vicent Sos Baynat s/n, 12071 Castellón, Spain*

Abstract

Biomedical knowledge resources (KRs) are mainly expressed in English, and many applications using them suffer from the scarcity of knowledge in non-English languages. The goal of the present work is to take maximum profit from existing multilingual biomedical KRs lexicons to enrich their non-English counterparts. We propose to combine different automatic methods to generate pair-wise language alignments. More specifically, we use two well-known translation methods (GIZA++ and Moses), and we propose a new ad-hoc method specially devised for multilingual KRs. Then, resulting alignments are used to transfer semantics between KRs across their languages. Transference quality is ensured by checking the semantic coherence of the generated alignments. Experiments have been carried out over the Spanish, French and German UMLS Metathesaurus counterparts. As a result, the enriched Spanish KR can grow up to 1,514,217 concepts (originally 286,659), the French KR up to 1,104,968 concepts (originally 83,119), and the German KR up to 1,136,020 concepts (originally 86,842).

Keywords: Semantic transference, multilingual biomedical knowledge resources, term alignment.

1. Introduction

Large-scale processing of textual data (e.g., scientific literature) has benefited from the extensive use of semantics described in biomedical knowl-

*Corresponding author. Tel.: +34 964 728370; Fax: +34 964 728435

Email addresses: mcatalan@uji.es (María Pérez), berlanga@uji.es (Rafael Berlanga)

edge resources (KRs) [1]. Semantics has been used for coding, indexing and retrieving domain-specific corpora. In the biomedical domain, most KRs are mainly expressed in English, like the Unified Medical Language System[1] (UMLS®) [2] and the NCBO BioPortal[2] [3]. In the case of UMLS, despite being multilingual, most information is expressed only in English. For example, the French projection of UMLS only covers the 7.5% of UMLS in English [4]. Therefore, applications that deal with information written in a non-English language suffer from this scarcity of knowledge. For example, hospital information systems usually require to align patients reports content with other biomedical data resources, but this implies to deal with different languages [5]. Another example is the need of multilingual annotated biomedical corpora to process knowledge as stated in [6] and [7]. Recently, many initiatives have aimed to provide non-English versions of these KRs. However, the translation gaps between English and other languages still remain large.

Automatic language translation has been largely used in the literature with the aim of translating English resources into other languages. However, as stated in [8], there are several issues that are hard to be considered by automatic approaches, e.g., the knowledge about the domain or linguistic issues such as regularities, genres, etc. Therefore, current automatic translation approaches suffer from ambiguity and lack of adequacy to specific domains.

The enrichment of KRs, in contrast to translation tasks, does not require exact lexical translations between languages, but word alignments between lexicons through which semantics are transferred. Word alignment has been used in many applications of natural language processing (NLP), namely: as a starting point of statistical translation (e.g., [9]), in cross-lingual information retrieval (e.g., [10]), in cross-lingual syntactic learning (e.g., [11, 12]), in word sense disambiguation (e.g., [13, 14]), and ontology matching (e.g., [15]).

In this paper we propose an automatic method to semantically enrich multilingual biomedical KRs through the use of implicit alignments present at these KRs. Here, we propose a new term alignment method that, in contrast to other approaches in the literature, considers statistic, lexical and semantic information. Our approach relies on the fact that biomedical terms are highly coordinated, i.e., most clinical terms are combinations of other

[1] http://www.nlm.nih.gov/research/umls

[2] http://bioportal.bioontology.org

elements described in the same thesaurus. This property is independent of the language, therefore we aim at identifying the implicit multilingual alignments that can be derived from these coordinated terms.

The outline of the paper is as follows. In Section 2, we review some approaches that aim to transfer data between resources expressed in different languages. Then, in Section 3, we describe general concepts that are used later in the description of the method. Section 4 describes the proposed approach. In Section 5, we evaluate the proposed approach and we show the results. Finally, Section 6 presents the main conclusions and future work. Results concerning aligned terms are publicly available at `http://krono.act.uji.es/STEM-KR`.

2. Related work

In the biomedical domain, there are several approaches that have addressed the need of transferring data between existing resources. Most of them aim to translate existing biomedical terminologies in English to non-English languages. These approaches can be classified depending on the technique used to make the translation, namely: morphological, corpus-based, and knowledge-based methods.

Morphological-based methods are focused on applying morphological information to derive word translation of medical terms. For example, [16, 17] build multilingual dictionaries using morphological relations. These approaches are language-dependent and they do not consider information about the domain.

Corpus-based methods perform word alignment using parallel corpora. Among them, there are methods that rely on existing parallel corpora. For example, [18, 19] used several parallel terminologies to build an English-Swedish dictionary. Other methods build their own parallel corpora to perform the alignment. For example, [20, 21] build a parallel corpora using web documents to find English-French translations of medical terms, [22] also searches English-French translations but in comparable corpora (i.e., text corpora addressing the same general topic in two different languages), and [23] uses a statistical vector model to match English UMLS terms with their German translations in a corpus aligned at document level. [24] uses parallel and comparable corpora to create an English-German bilingual lexicon and to enrich multilingual thesauri. The proposed method uses morphological and statistical information to get the terms alignments, but its coverage is quite

poor. Recently, the CLEF-ER challenge [25] has concerned with the multilingual alignment of biomedical document corpora. More specifically, this challenge proposed to recognize biomedical entities in three parallel corpora (MedLine, EMEA and Patents), as well as to report term correspondences between language pairs. Evaluation was performed against a silver standard corpus (SSC), which was built from the annotation agreements of the participants. Our work differs from this challenge in that we aim at enriching non-English KRs by identifying the implicit alignments present at the parallel KRs, instead of looking for them in parallel document corpora. Dealing with the KR lexical information allows us to get higher quality term alignments as well as a better coverage of the different semantic types of the KRs. The resulting enriched KRs could be indeed used to perform the tasks proposed in CLEF-ER.

Knowledge-based methods use the data stored in the KRs to perform the alignments rather than a parallel corpus. For example, [21] uses the UMLS Concept Unique Identifiers (CUIs) to integrate information from various terminologies, considering in this way, synonyms and translations in other languages (whenever the CUI is available in the non-English terminologies). At the end, corpus-based alignments are combined with the alignments retrieved through the CUIs. [26] combines the knowledge stored in UMLS with lexical information in order to translate the Foundational Model of Anatomy (FMA) ontology into French.

Most of these approaches first apply an alignment algorithm and, later, filter the results to get only translations of terms valid in the domain. For example, [27] selects only the terms relevant to the domain by checking their occurrence in biomedical corpora, and [21] selects the biomedical terms by checking if they appear in biomedical terminologies.

In this paper, we propose a corpus and knowledge-based automatic approach that combines statistical, lexical and semantic information to perform term alignment. At the end, a semantic filtering is performed to select alignments that are semantically coherent within the domain. In contrast to other approaches in the literature, we use semantics to select the relevant biomedical alignments.

3. Background

In this section we introduce the concepts and foundations that underlie the proposed method. First, we define the concept of semantics used in this

work, and how semantics are expressed in current KRs. Then, we define the process of transference of semantics and, finally, we introduce the notion of term alignment as the main foundation of our approach.

3.1. Knowledge resources

From a broad perspective, the concept *semantics* refers to the study of the meaning. It relates words, phrases and symbols with their meaning, which implies relations between concepts and categorization, among other issues.

A knowledge resource (KR) is a formalization of the semantics of a domain by means of a set of concepts which represents meaningful entities of the domain, and a set of relations between them.

A concept is usually characterized by: *(i)* a concept identifier, *(ii)* a set of labels that includes synonyms and short descriptions, which can be terms or even sentences, and *(iii)* a definition or gloss. The set of labels contains the strings describing the concept, and also lexical variants of these strings.

Moreover, concepts can be taxonomically related by subsumption (*is-a*) or "broader-than" relationships. More formal KRs also define logical axioms between concepts, e.g., OWL ontologies, such as the National Cancer Institute Thesaurus (NCI) and FMA.

Usually, the domain covered by a KR is divided into a set of subdomains (or categories) that have specific characteristics. These subdomains can be partially ordered by the subsumption relationships. In this paper, we adopt the semantic groups of UMLS [28] to define these subdomains.

3.2. Semantic transference

Semantic transference refers to the assignment of semantics to terms that are not yet described in a target KR, by considering the information available of these terms in a source KR. In this work, we deal with the transference of semantics across languages within multilingual KRs.

More specifically, a multilingual KR is a KR in which the labels describing a concept are expressed in different languages. KR^{lang} is the projection of the KR to the language *lang*.

Unfortunately, there is usually a large difference between the coverage of languages in existing multilingual KRs. For example, in UMLS, as stated in [4], non-English counterparts lack between 65% to 94% of the coverage of the English UMLS. Particularly, the Spanish projection only covers the 35% of the English UMLS vocabulary, whereas French only achieves the 7.5%.

3.3. Term alignment

In this paper we propose a language alignment method to perform the transference of semantics between languages. Language alignment can be done at different levels, from document alignment to term alignment, with paragraph and sentence alignment in between.

The most popular word alignment techniques are the Hidden Markov Models (HMMs) [29] and the IBM models [30, 31]. However, the most sophisticated IBM models only achieve to get many-to-one mappings, while real word alignments have many-to-many mappings (i.e., one token in the source language can correspond to multiple tokens in the target language, and the opposite). Some approaches propose combinations or modifications of the IBM models in order to achieve many-to-many mappings, e.g., [32] performs the intersection of IBM models bidirectional alignments, and [34, 35] combine symmetrization with a maximum entropy approach.

In this work, we perform term alignment to transfer the semantics of the terms described in a source KR^{l1} to a target KR^{l2}, where *l1* is usually English and *l2* is another language.

Given a source text and a target text consisting of word sequences, a term alignment is a correspondence between subsequences of words in the source text and subsequences of words in the target text.

We consider terms (sequences of words) instead of single words, because there are sequences of words that have a different meaning from the meaning of its individual words [36], for example, *cauda equina* and *cáscara sagrada*.

In this work, we evaluate the correctness of term alignments by analyzing their *semantic coherence*. Two aligned terms must have similar semantics since they are supposed to be equivalent in their respective languages. Therefore, we assume that an alignment is semantically coherent if the source term and the target term have similar semantics in their respective KR projections.

4. Materials and methods

In this section, we describe the proposed approach to automatically enrich multilingual KRs. First, it finds out pair-wise language alignments and, then, these alignments are used to transfer semantics between KRs. Figure 1 shows the overview of the approach.

To obtain the language alignments, we propose an alignment method based on statistical, lexical, and semantic information in order to get the

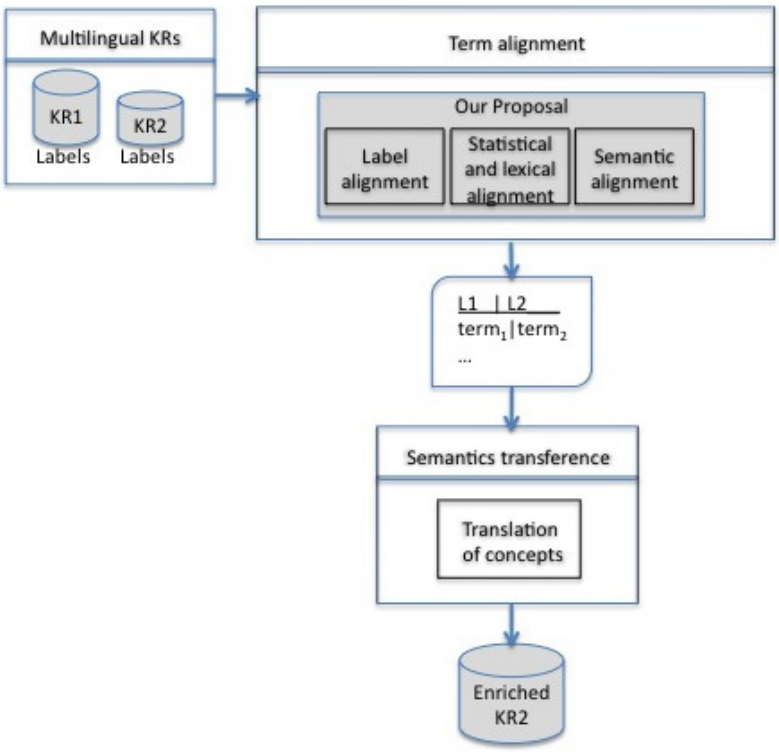

Figure 1: Overview of the semantics transference approach.

maximum coverage of alignments. Thus, the method requires a parallel corpus where to learn these alignments. In our approach, this parallel corpus is directly derived from the KR lexicons.

The most remarkable features of the proposed alignment method are summarized here:

- It finds out not only one-to-one word alignments, but also many-to-many words alignments. From now on, we refer to these alignments as *term alignments*.
- It combines statistical and lexical information to find out the most likely alignment, and semantic information to find out synonyms.
- It finds out alignments of words that may not be described in the KRs.

- It tailors the alignments to best fit the domain by using the semantics covered by the KRs.

The last step is the transference of semantics, which consists in translating labels from the source KR into the target KR by using the generated alignments.

In this paper, we deal with English, Spanish, French, and German languages, but the approach is language-independent as it does not rely on any linguistic tool such as POS-tagging, or syntactic analysis.

In next sections, the term alignment and the semantic transference methods are further described. Previously, we describe how the parallel corpora required by the method is built from the lexicons of the multilingual KRs.

4.1. Creation of the parallel corpora

The first task consists in creating the parallel corpora that contain alignments of sentences in the languages of interest. In this paper, we build these parallel corpora with the lexicons and knowledge provided by the KRs. More specifically, for UMLS we have used the MRCONSO file[3], which accounts for all labels associated to the KR concepts. To build a parallel corpus, we select the labels of the concepts that are described in both languages in the KR. It is important to remark that the resulting alignments are not sentence to sentence alignments, since a concept can have several labels in a same language. In consequence, a KR alignment derived from concept c has the following structure:

$$d_1, d_2, ..., d_n \overset{c}{\leftrightarrow} d'_1, d'_2, ..., d'_m$$

where each d_i is a label formed by a sequence of words $(w_1 w_2 ... w_k)$. Notice that KR alignments are derived from the labels associated to a same concept expressed in several languages. Thus, labels d_i are expressed in the source language, and d'_j in the target language.

An example of an entry with multiple labels in the EN-ES parallel corpus is the following:
"Amnioscopy", "obstetrics endoscopy" $\leftrightarrow$ *"amnioscopia", "endoscopia obstétrica"*

[3]http://www.ncbi.nlm.nih.gov/books/NBK9685/

4.2. Term alignment

The goal of this step is to find out correspondences of a term of the source language in the target language. These correspondences can be identified through the analysis of the alignments of the labels in the parallel corpus.

Unfortunately, the correspondence between labels is not direct, and we need to identify which labels are equivalent or similar to properly identify the term alignments. Therefore, the alignment of languages is divided into two steps: *(i)* alignments of labels, and *(ii)* alignments of terms.

4.2.1. Alignment of labels

An entry in the parallel corpus describes a many-to-many correspondence between labels. In case there is only one label in each language, the correspondence is direct. Otherwise, to determine the best correspondences between labels, we calculate the following probability that uses a word entailment model built on the parallel corpus:

$$P(d'|d) = \prod_{w \in d, w' \in d'} P(w'|w) \tag{1}$$

With this model, we obtain pairs (d, d'), where d' is the label in the target language that maximizes the probability for d. $P(w'|w)$ is the probability of seeing w' in the target language when we have seen w in the source language. This probability can be easily estimated with maximum likelihood estimation (MLE) from the parallel corpora and smoothed with some simple method (e.g., Laplace) to avoid zero probabilities.

In the example shown above, the labels alignments are:

amnioscopy - amnioscopia

obstetrics endoscopy- endoscopia obstétrica

4.2.2. Alignment of terms

Given a pair of aligned labels (d, d'), the next step is to find out alignments between the terms in those labels. First, we look for alignments of terms considering the statistics and lexical characteristics of the labels in which they appear. Then, we refine the alignment set by using the KR semantics (i.e., semantic groups). In next sections, both techniques are explained.

Step 1: Statistical and lexical alignments

The first step of the term alignment takes into account statistical information and the characteristics of the context in which the terms appear, that is, the labels. We use an iterative algorithm, described in Algorithm 1, that identifies terms alignments by searching subsumed alignments within the terms. This algorithm iterates on the number of unigrams, n, that compose the terms. First, it checks whether the source term can be translated or not into the target term by applying the available alignments. If it cannot be translated and it has n-grams, $d - d'$ is considered a new alignment. Then, it searches for implicit alignments within the terms, removing those words that have been already aligned. For each non-aligned word, it searches for the most likely target word, which is selected if the score of the alignment is higher than zero. The score of an alignment is given by the function:

$$score(w, w') = \alpha \cdot P(w'|w) + (1 - \alpha) \cdot similarity(w, w') \tag{2}$$

$P(w'|w)$ is the probability of seeing w' in the target language when we have seen w in the source language, and the function $similarity(w, w')$ measures the lexical similarity between the aligned terms (Levenshtein's distance). We introduce lexical similarity as a means to favour the mapping of similar tokens when their probabilities are too low. In the experiments, the weight of the lexical component plays a minor role, being set to 0.2 independently of the language pair.

The result of this algorithm is a set of alignments in which words or sequences of words in the source language are aligned to words or sequences of words in the target language. Multiple alignments are possible since the algorithm can identify different valid alignments in different aligned labels.

Step 2: Semantic alignments

The purpose of this step is to refine the resulting alignments by considering the KR semantics. Given an alignment (t, t'), we perform a dictionary look-up to match t to concepts from the KR. For each concept c matching the term t, we store t' in order to relate also the translations of synonyms of t (terms also annotated with the concept c). Therefore, when two different terms t_1 and t_2 are annotated with the same concept, they will share the translations $[t'_1, t'_2]$. Then, to tailor the set of alignments in order to best fit the domain, we select only those alignments that are semantically represented in the target KR. Algorithm 2 describes the semantic alignment of terms.

Algorithm 1 Alignment of terms.

procedure ALIGN_TERMS(d,d',*alignments*,n)
if d can be translated to d' by the available alignments **then return** alignments
end if
if $|d| = n$ **then**
append (d, d') to *alignments*
end if
$d^* \leftarrow$ remove from d those terms $t \in d$ that appear in *alignments* as (t, t') with $t' \in d'$
$d'^* \leftarrow$ remove from d' those terms $t' \in d'$ that appear in *alignments* as (t, t') with $t \in d$
for w in d^* **do**
for w' in d'^* **do**
$score[w'] = \alpha \cdot P(w'|w) + (1 - \alpha) \cdot similarity(w, w')$
end for
$w'_{max} \leftarrow argmax(score)$
if $score[w'_{max}] > 0$ **then**
append (w, w'_{max}) to *alignments*
end if
end for
return $Align\_terms(d, d', alignments, n + 1)$
end procedure

Algorithm 2 Semantic alignment of terms.

procedure SEMANTIC_ALIGNMENT($alignments, KR_2$)
  $new\_alignments = \{\}$
  $conceptual\_alignment = \{\}$
  for (t, t') in $alignments$ **do** ▷ $Step1.\ Synonyms$
    $c \leftarrow semantic\_annotation(t)$
    append t' to $conceptual\_alignment[c]$
  end for
  for (t, t') in $alignments$ **do** ▷ $Step2.\ Domain filtering$
    append $t' \cap KR_2[c]$ to $new\_alignments[t]$
    if $new\_alignment[t] = []$ **then**
      append $conceptual\_alignment[c] \cap KR_2[c]$ to $new\_alignments[t]$
    end if
    if $new\_alignment[t] = []$ **then**
      append t' to $new\_alignments[t]$
    end if
  end for
return new_alignments
end procedure

In the previous example, the obtained alignments are:

amnioscopy -amnioscopia

obstetrics -obstétrica

endoscopy-endoscopia

Analyzing the results of each step of the alignment method, the use of semantics includes new alignments by synonymy and rejects out-of-domain alignments. For example, *amenia* is aligned to *falta de menstruación* with statistical and lexical information. By the use of semantics it is also aligned to *amenia* and *amenorrhoea*. An example of rejected alignment is *junk-chatarra*, which is not specific of the domain, whereas *junk-heroína* and *junk-diamorfina* remain in the alignment set.

4.3. Transference of semantics

Once the set of alignments is generated, the last step consists in transferring semantics between KRs. In this paper, we transfer semantics by applying

a simple translation of source labels along with their CUIs to the target KR, whenever these CUIs are not present in the latter.

More specifically, given a source label l, we look for the longest subsequences of words in l that appear as source terms in at least one alignment (we take the top ranked alignment considering the score in Equation 2).

If all words of the label of a CUI can be translated, then the CUI is transferred and a new entry with the CUI and the translated label is added to the target KR. Finally, target language constraints expressed as word entailment distributions ($P(w|w')$) are applied in order to select the most appropriate variants as well as word ordering for the translated label.

5. Results

In this section, we show the results of the experiments carried out to evaluate the proposed semantic transference approach. We have performed the experiments over UMLS (version 2012AB) in order to enrich their Spanish, French and German counterparts. For this purpose, we have executed our alignment method to the pairs of languages shown in Figure 2. To evaluate the correctness of the resulting alignments, we have performed two experiments, one based on semantic information and another one based on the validation through an external reference dataset:

- **Semantic coherence evaluation**. We assume that in a correct alignment, the source term and the target term must have similar semantics. As earlier mentioned, we describe the semantics of a word with the semantic groups of UMLS Semantic Network[4].

  More formally, given a multilingual alignment $a = \{t_1, t_2\}$, its semantic coherence is the semantic overlap of its terms, namely:

  $$semantic\_coherence(\{t_1, t_2\}) = |sem\_group(t_1) \cap sem\_group(t_2)| \quad (3)$$

  where $sem\_group(t)$ returns the set of UMLS semantic groups of the CUIs having t as label. We consider an alignment semantically coherent when the semantic overlap of its terms is greater than zero.

  The evaluation of the semantic coherence is automatically performed over the system-generated alignments by applying the previous formula.

[4]http://semanticnetwork.nlm.nih.gov/SemGroups/SemGroups.txt

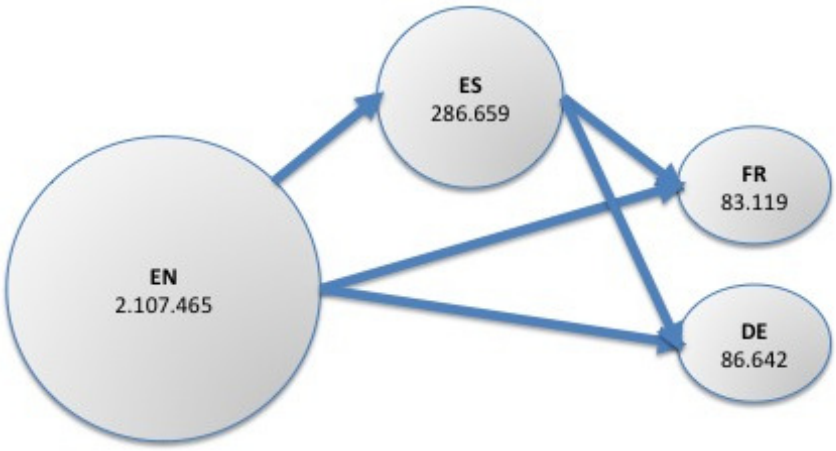

Figure 2: Number of concepts and transference between multilingual parallel KRs.

Quality of alignments are then evaluated by counting the number of semantically coherent alignments each method generates.

- **External reference validation**. Alignments can also be validated by using an external resource as reference. In this experiment, we have used BabelNet [5] (version 2.0), a multilingual dictionary plus a semantic network over it. Notice that we use explicit BabelNet alignments as an indirect indicator of the quality of the alignments found by each method.

Finally, we evaluate the semantic transference by analyzing the number of concepts successfully translated, and by estimating their precision over a small subset of translated labels.

In next sections, we describe the different experiments we have carried out to evaluate and validate our approach. First, in Section 5.1, we describe the main characteristics of the parallel corpora used in the term alignment. Section 5.2 shows the results of the execution of the alignment method and the evaluation of the resulting alignments by comparing them with two well-known statistical-based alignments methods, GIZA++ [31] and Moses [9], two of the most popular alignment algorithms. While GIZA++ only performs unigram alignments, Moses performs multiwords alignments. Finally,

[5]http://babelnet.org

	EN-ES	EN-FR	ES-FR	EN-DE	ES-DE
CUIs in common	260,961	78,127	69,930	78,458	72,578
Avg. labels in L1	2.5	3.74	2.38	3.71	2.36
Avg. labels in L2	1.89	1.98	1.98	1.92	1.92
1 label - 1 label	87,566	21,115	26,429	21,629	28,773
Unigram - Unigram	5,951	2,071	4,992	2,431	5,737
Unigram - Ngram	692	188	150	108	90

Table 1: Characteristics of the parallel corpora (ENglish, FRench, ES-Spanish and DE-German.)

in Section 5.3, we present the results of the transference of semantics between KRs.

5.1. Preparation of the parallel corpora

For each pair of languages shown in Figure 2, we have created a parallel corpus by selecting those concepts in UMLS that have labels in both languages. Table 1 shows the main characteristics of each corpus. For example, in EN-ES parallel corpus, there are 260,961 concepts in common, with an average of 2.5 labels per concept in EN and 1.98 labels per concept in ES. Therefore, the EN-ES parallel corpus has 260,961 entries, in which 87,566 are one label - one label, of which 6,633 are explicit word alignments (5,951 unigram-unigram and 682 unigram-multiword).

5.2. Term alignment results

The results of the term alignment method for the different language pairs are shown in Figure 3. The alignments are classified by the number of words that compose the source term. In all cases, the number of unigrams is high, and if we compare these numbers with the explicit unigram alignments (those in the parallel corpus), we can conclude that our method is able to find a considerable amount of implicit alignments, whose correctness is evaluated in the following sections.

5.2.1. Unigram alignments evaluation

In this section, we compare the unigram alignments obtained with our approach against the results of executing GIZA++ with the parallel corpora

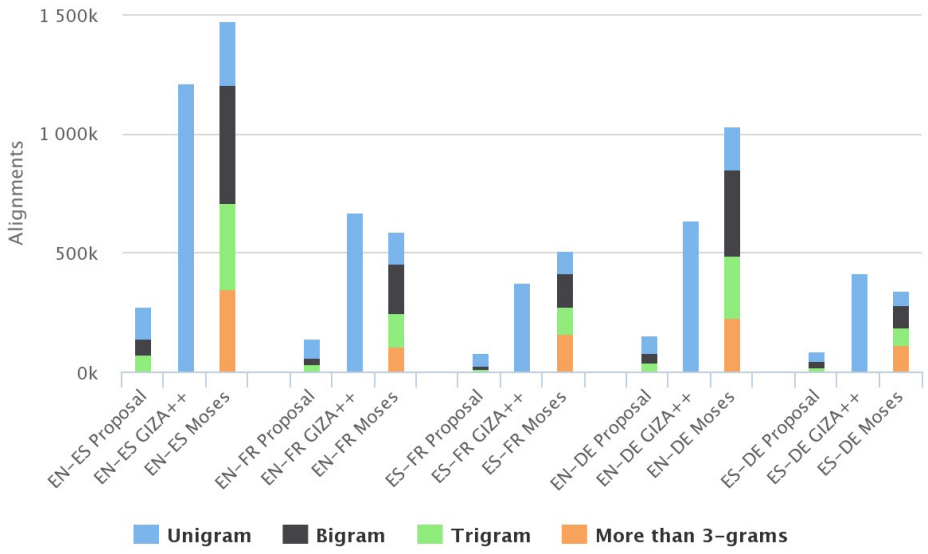

Figure 3: Alignments of our approach, GIZA++ and Moses with the parallel corpora.

described above. Figure 3 shows the results of executing GIZA++ with each one of the parallel corpora.

GIZA++ provides an average of six alignments per source term, and to select the most relevant alignments for a source term, we have defined a filtering strategy. First, we select the top-k alignments whose probability is higher than a threshold λ_{prob}, and the difference between the probability and the maximum probability for the source term is lower than a threshold β. Then, we evaluate the selected alignments with the function defined in Equation 2. If the returned value is higher than a threshold $\lambda_{f(a)}$, the alignment is selected. For example, the word *ileoscopy* is originally aligned by GIZA++ to {*endoscopia, examen, endoscópica, operación, íleon, ileoscopia, ileoscopy, fibroileoscopia,...*}. After filtering, only the alignment to *ileoscopia* is selected, which is the correct one.

To compare the results of GIZA++ and those of our approach, we analyze the semantic coherence of the alignments. We have parameterized the filtering to get similar sets of alignments (in size) to the ones of our approach in order to compare them. We select only the unigrams of our approach since GIZA++ aligns only uniwords. Figure 4 shows the evaluation of the semantic coherence of the alignments of GIZA++ and the unigram alignments of

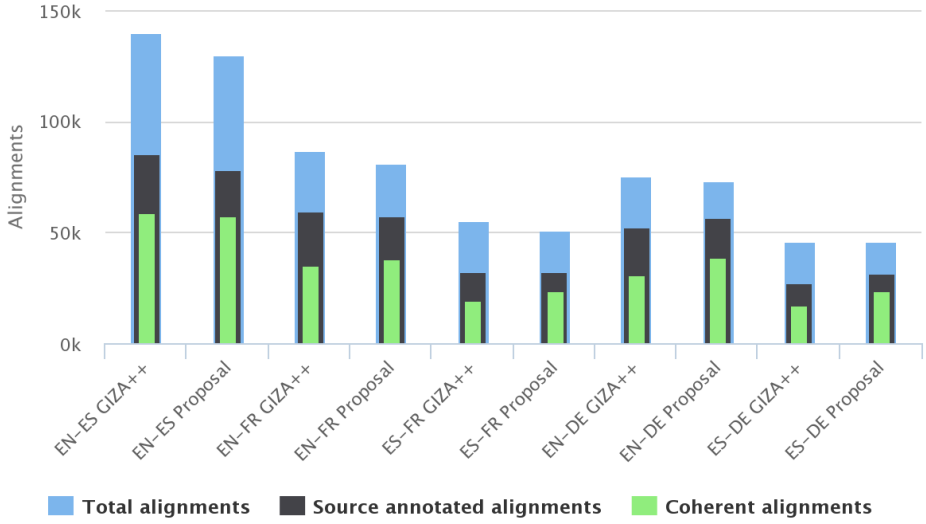

Figure 4: Semantic coherence evaluation of GIZA++ filtered alignments and the unigram alignments of our approach in which the source terms are annotated.

our approach. It shows the total number of alignments, the number of alignments whose source term has a CUI associated, and the number of semantically coherent alignments. Examples of GIZA++ incoherent alignments are: $abstracting^{ACTI} - resumen^{CONC}$ and $aeration^{PROC} - ventilación^{DISO}$.

In all cases, the number of analyzed alignments is similar, but the semantic coherence is slightly better in our approach. However, as GIZA++ does not consider multiword terms, it often generates wrong one-to-one alignments when they are actually one-to-many. For example, the word *achromic* is aligned by GIZA++ to *pigmentado*, but the correct alignment is *no pigmentado*.

5.2.2. Term alignments evaluation

In this section we evaluate the complete set of alignments by comparing them to the results obtained by Moses. In contrast to GIZA++, Moses considers also multiword terms in its language models. Examples of Moses multiwords alignments (together with their conditional probability) are the following ones:

fluid granulocyte count - *recuento de granulocitos en líquido* (0.5)

fluid granulocyte count - *cuenta granulocitaria en* (0.5)

	EN-ES	**EN-FR**	**ES-FR**	**EN-DE**	**ES-DE**
Moses Original set	406,708 (27.61%)	130,092 (22.17%)	152,929 (29.92%)	157,496 (16.92%)	45,990 (16.25%)
Moses Filtered	35,188 (12.59%)	15,226 (13.67%)	18,272 (18.09%)	10,807 (4.13%)	7,748 (8.59%)

Table 2: Number of redundant alignments and the percentage w.r.t. the original and filtered sets (1,2,3-grams).

performance at - desempeño en el (1.0)

hand application of - colocación de un (0.5)

hand application of - colocación de (0.5)

These examples show that some alignments only differ in one meaningless word. For example, the two alignments of the source term *hand application of* are indeed equivalent. This is due to the fact that Moses considers all words in the same way independently of their nature.

Figure 3 shows the number of alignments found out by Moses for each parallel corpus. As it can be seen, most alignments are unigrams, bigrams and trigrams. Since biomedical terminologies are highly coordinated, longer terms can be usually decomposed into smaller terms. As a consequence, most long terms are indeed redundant for performing semantics transference.

We have carried out an experiment to measure the redundancy of Moses alignments. Table 2 shows the number of alignments that are redundant in the original set of Moses alignments and in a filtered set comparable to our set of alignments. This filtered set is the result of applying the filtering strategy described in Section 5.2.1 to 1,2,3-grams. As it can be seen, original sets exhibit a high level of redundancy, which is notably reduced when only 1,2,3-grams are selected. An example of redundant alignment is *blood selenium-selenio en sangre*, which can be decomposed and translated by its individual terms *blood-sangre* and *selenium-selenio*. Another example is *induction of labour-inducción del trabajo de parto*, which can be decomposed into *induction-inducción* and *labour-trabajo de parto*

Figure 5 compares the Moses filtered alignments sets with those generated by our proposal. Although the number of filtered Moses alignments is

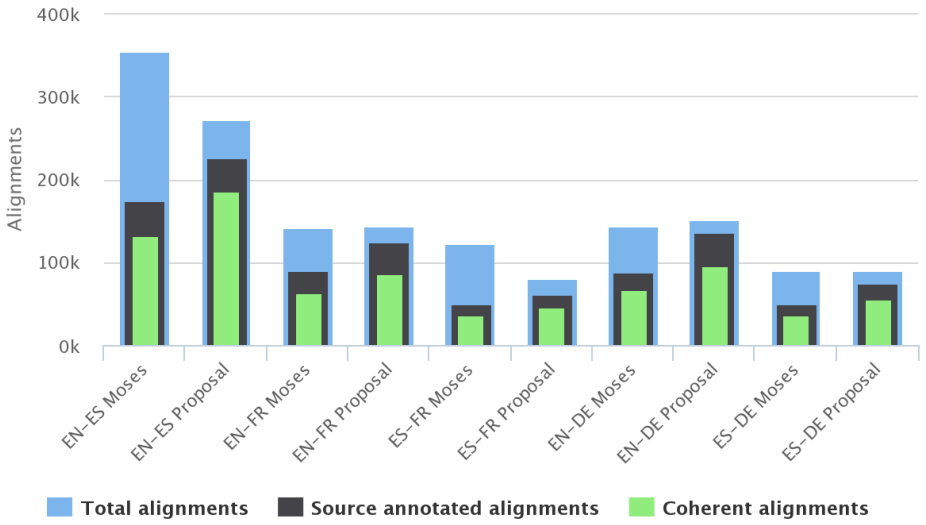

Figure 5: Semantic coherence evaluation of Moses filtered alignments and the alignments of our approach in which the source terms are annotated.

higher than ours, the number of annotated source words is higher in our approach. That is, Moses alignments contains more source words that are not semantically described in the KR. Our approach obtains more semantically coherent alignments than Moses in all language pairs.

5.2.3. Validation of the alignments using an external resource

In this section we present the results of the validation of the correctness of the alignments using BabelNet 2.0 as external reference. BabelNet defines the entries with unique identifiers which allow us to align concepts in different languages. The English version of BabelNet 2.0 has 932,596 concepts (excluding named entities), while the Spanish has 425,914 concepts, the French version has 256,813 concepts, and the German version has 220,136 concepts. We want to remark that BabelNet is a general domain KR and we cannot consider it as a GS, since it scarcely covers the bioinformatics domain. Nevertheless, we consider relevant the use of BabelNet to identify alignments of some entities not so well covered by UMLS (e.g., geographical terms).

In this experiment, we have validated the alignments of our approach, GIZA++ and Moses. Table 3 shows the number of alignments of the three approaches that are explicitly set in BabelNet, and the coverage of these

	EN-ES	**EN-FR**	**ES-FR**	**EN-DE**	**EN-DE**
Our Proposal	32,696 (1.12%)	23,225 (0.93%)	14,603 (1.13%)	22,695 (0.65%)	13,709 (0.84%)
GIZA++ Filtered	21,909 (0.7%)	16,821 (0.7%)	9,419 (0.73%)	11,882 (0.34%)	6894 (0.4%)
Moses Filtered	21,407 (0.73%)	13,396 (0.54%)	10,007 (0.77%)	22,480 (0.65%)	8,266 (0.5%)

Table 3: Number of alignments that appear in BabelNet and the coverage of these alignments with respect to BabelNet alignments.

UMLS-derived alignments. As expected, the low overlapping is due to the difference in the domains BabelNet and UMLS are focused on. Comparing the alignments of our approach and the filtered sets of GIZA++ and Moses, our proposal obtains the highest number of shared alignments with BabelNet in all the languages pairs.

5.3. Semantic transference evaluation

The last step of our method is the transference of semantics between KRs through the selected alignments. In this experiment, we transfer semantics from richer to poorer covered KRs (see Figure 2). We only consider semantically coherent alignments in order to ensure the correctness of the translation. Figure 6 shows the size of the enriched KRs by the alignments of our approach, and the filtered alignments of GIZA++ and Moses. It shows the number of concepts whose labels have been translated, and the number of invariant concepts which do not need translation. These invariant concepts correspond to named entities and latin expressions (e.g., species scientific names).

We have also joined the alignments of the three approaches (*Combination*), which increases the number of translated concepts with respect to our method. This shows that the three approaches complement each other and, therefore, the coverage of the alignments is higher. As result of this combination, in the enriched Spanish KR, the 41% of the concepts have been transferred by translation, the 40.8% are invariant concepts, and 18.2% appear in the original Spanish KR. However, as expected, in the French and German counterparts, the transference of concepts by translation is poorer than in Spanish due to the size of their KR lexicons. In fact, the percentages of the translated concepts are 36.9% and 37.7% in French and German

Lang.	ANAT	DISO	CHEM	PROC	PHYS	OBSV	Total
ES	0.9	0.9	0.9	0.9	0.7	0.8	0.85
FR	0.9	1.0	0.8	0.8	0.9	0.8	0.87
DE	0.8	0.7	0.9	1.0	0.7	0.7	0.8

Table 4: Precision of the translations of the concepts of the most frequent semantic groups.

respectively, and the percentages of invariant concepts are 55.9% and 54.9%, respectively.

Moreover, we have also included BabelNet alignments in order to see if BabelNet covers the vocabulary that is not aligned by the three approaches. However, as Figure 6 shows, the improvement in the semantics transference is low. Most of the words that are not translated correspond to chemical products e.g., *profollipsin, aminopolypeptidase, chloroestradien*, acronyms e.g., *URP, CSRP, SMF*, and protein names, e.g., *chordin, RHCE*, among others.

It is also worth mentioning that transferred concepts in French and German KRs, 90% come from English and 10% from the Spanish KR.

Finally, to evaluate the quality of the transference, we have performed a manual validation on a subset of translated labels. We have randomly selected 50 translations per language from concepts of the most frequent semantic groups, and we have evaluated the precision of these translations (i.e., number of correct translations w.r.t. the evaluated translations). The translations and their evaluation are publicly available at `http://krono.act.uji.es/STEM-KR`. It is worth mentioning that we consider the translations as bag of terms. A translation is considered correct if all the component terms are correct translations of the original label in the context they are expressed.

Table 4 shows the precision of the translations in each language grouped by semantic group. As it can be shown, the German translations subset has a poorer precision than the subsets of the other two languages. The main reason is the intrinsic lexical characteristics of the German language, in which terms are compound words instead of coordinate expressions. As future work, we aim to decompose German words in order to obtain more accurate alignments.

5.4. Comparison with other approaches

With respect to other approaches that also aim to enrich multilingual biomedical KRs, our proposal clearly outperforms them in scale. For exam-

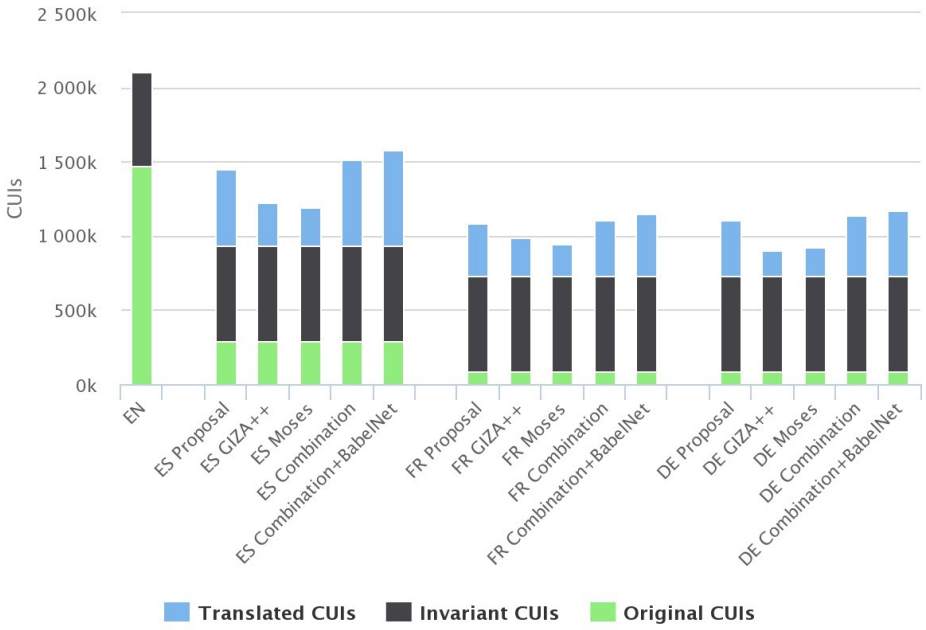

Figure 6: Size of the enriched KRs after semantic transference.

ple, [24], which aims to enrich a multilingual thesaurus (specifically German) and to create a bilingual lexicon, only adds 1400 new German terms to the German KR. [26] addresses the translation of FMA using knowledge-based and lexical methods. Using UMLS as KR, they translate 7,469 concepts into French, and using the lexical approach they get 6,246 correspondences between English and French. Therefore, our approach performs a considerable larger transference than these approaches.

6. Conclusions

In this paper we have dealt with the problem of the scarcity of semantic knowledge in non-English languages in the biomedical domain. We have proposed an automatic term alignment method to transfer semantics in multilingual knowledge resources.

The results of the experiments show that the proposed method is able to find out implicit alignments in a multilingual KR, with which the semantic transference between English and other languages can be automatically performed. We have compared our approach with two of the most used alignments methods, GIZA++ and Moses, and our method gets more semantically coherent alignments. However, we have seen that the three approaches complement each other, and their combination increases considerably the transference of semantics. As a result of this combination, the Spanish UMLS counterpart grows up to 1,514,217 concepts, the French counterpart up to 1,104,968 concepts, and the German counterpart up to 1,136,020 concepts.

As future work, there are several interesting research lines derived from this work. First, we are going to address the decomposition of German terms in order to get more accurate alignments. Moreover, we will study how to translate the non-covered entities by taking into account external corpora specific to this domain. We also plan to use the semantic relationships defined in the KR to further enrich and improve the coherence in the different language KR counterparts.

Acknowledgements

We thank anonymous reviewers for their very useful comments and suggestions. The work was supported by the R&D project TIN2014-55335-R from the Spanish Ministry of Economy and Competitiveness (MINECO) and by the UJI INNOVA project 13I346 in collaboration with ActualMed.

References

[1] K.-H. Cheung, E. Prudhommeaux, Y. Wang, S. Stephens, Semantic Web for Health Care and Life Sciences: a review of the state of the art, Briefings in Bioinformatics 10 (2) (2009) 111–113.

[2] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (suppl 1) (2004) D267–D270.

[3] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications., Nucleic Acid Research 39 (2011) W541–W545.

[4] J. Hellrich, U. Hahn, Enhancing Multilingual Biomedical Terminologies via Machine Translation from Parallel Corpora, in: Natural Language Processing and Information Systems, vol. 8455 of *Lecture Notes in Computer Science*, Springer International Publishing, ISBN 978-3-319-07982-0, 9–20, 2014.

[5] A. Roberts, R. Gaizauskas, M. Hepple, N. Davis, G. Demetriou, Y. Guo, J. S. Kola, I. Roberts, A. Setzer, A. Tapuria, B. Wheeldin, The CLEF corpus: semantic annotation of clinical text., in: AMIA Annual Symposium Proceedings, vol. 2007, 625–629, 2007.

[6] R. Berlanga, A. Jimeno-Yepes, M. Pérez-Catalán, D. Rebholz-Schuhmann, Context-Dependent Semantic Annotation in Cross-Lingual Biomedical Resources, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, vol. 8138 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, ISBN 978-3-642-40801-4, 120–123, 2013.

[7] L. Deléger, A.-L. Ligozat, C. Grouin, P. Zweigenbaum, A. Neveol, Annotation of Specialized Corpora using a Comprehensive Entity and Relation Scheme, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), ISBN 978-2-9517408-8-4, 2014.

[8] C. Schaffner, Running before walking? Designing a translation programme at undergraduate level., Developing translation competence (2000) 143–156.

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, 177–180, 2007.

[10] J. Wang, Matching meaning from cross-language information retrieval, Ph.D. thesis, University of Maryland, 2005.

[11] R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, O. Kolak, Bootstrapping Parsers via Syntactic Projection Across Parallel Texts, Nat. Lang. Eng. 11 (3) (2005) 311–325, ISSN 1351-3249.

[12] D. A. Smith, N. A. Smith, Bilingual parsing with factored estimation: Using English to parse Korean, in: In Proc. of EMNLP, 49–56, 2004.

[13] Y. S. Chan, H. T. Ng, Scaling Up Word Sense Disambiguation via Parallel Texts, in: Proceedings of the 20th National Conference on Artificial Intelligence, vol. 3 of *AAAI'05*, ISBN 1-57735-236-x, 1037–1042, 2005.

[14] H. T. Ng, B. Wang, Y. S. Chan, Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study, in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1 of *ACL '03*, 455–462, 2003.

[15] C. Meilicke, R. García-Castro, F. Freitas, W. R. van Hage, E. Montiel-Ponsoda, R. R. de Azevedo, H. Stuckenschmidt, O. vb Zamazal, V. Svtek, A. Tamilin, C. Trojahn, S. Wang, MultiFarm: A benchmark for multilingual ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web 15 (0) (2012) 62 – 68, ISSN 1570-8268.

[16] C. Lovis, R. Baud, A. Rassinoux, P. Michel, J. Scherrer, Medical dictionaries for patient encoding systems: a methodology, Artificial Intelligence in Medicine 14 (1998) 201 – 214, ISSN 0933-3657.

[17] V. Claveau, P. Zweigenbaum, Translating biomedical terms by inferring transducers, in: Proceedings of the 10th Conference on Artificial Intelligence in Medicine, 2005.

[18] M. Nystrom, M. Merkel, L. Ahrenberg, P. Zweigenbaum, H. Petersson, H. Ahlfeldt, Creating a medical English-Swedish dictionary using interactive word alignment, BMC Medical Informatics and Decision Making 6 (35).

[19] M. Nystrom, M. Merkel, H. Petersson, H. hlfeldt, Creating a medical dictionary using word alignment: The influence of sources and resources, BMC Medical Informatics and Decision Making 7 (1) 37.

[20] L. Deléger, M. Mergel, P. Zweigenbaum, Translating medical terminologies through word alignment in parallel text corpora, Journal of Biomedical Informatics 42 (4) (2009) 692–701.

[21] L. Deléger, T. Merabti, T. Lecrocq, M. Joubert, P. Sweigenbaum, S. Darmoni, A Twofold Strategy for Translating a Medical Terminology into French, in: AMIA Annual Symposium Proceedings, 2010.

[22] Y. Chiao, P. Zweigenbaum, Looking for French-English translations in comparable medical corpora, in: Proceedings of AMIA Symposium, 2002.

[23] D. Widdows, B. Dorow, C. ki Chan, Using Parallel Corpora to enrich Multilingual Lexical Resources, in: In Third International Conference on Language Resources and Evaluation, 240–245, 2002.

[24] H. Déjean, E. Gaussier, J. M. Renders, F. Sadat, Automatic Processing of Multilingual Medical Terminology: Applications to Thesaurus Enrichment and Cross-language Information Retrieval, Artif. Intell. Med. 33 (2) (2005) 111–124, ISSN 0933-3657.

[25] D. Rebholz-Schuhmann, S. Clematide, F. Rinaldi, S. Kafkas,E. van Mulligen,C. Bui,J. Hellrich,I. Lewin,D. Milward,M. Poprat,A. Jimeno-Yepes,U. Hahn,J. Kors, Wntity Recognition in Parallel Multi-lingual Biomedical Corpora: The CLEF-ER Laboratory Overview, Information Access Evaluation, Multilinguality, Multimodality, and Visualization, 8138, 353-367, 2013.

[26] T. Merabti, L. Soualmia, J. Grosjean, O. Palombi, J.-M. Mller, S. Darmoni, Translating the foundational model of anatomy into french using knowledge-based and lexical methods, BMC Medical Informatics and Decision Making 11 (1) 65, 2011.

[27] E. K. van Mulligen, Q.-C. Bui, J. A. Kors, Machine Translation of BioThesauri, in: Proceedings of the CLEF-ER Workshop 2013, 2013.

[28] A. T. McCray, A. Burgun, O. Bodenreider, Aggregating UMLS Semantic Types for Reducing Conceptual Complexity, Studies in Health Technology and Informatics 84 (0 1) (2001) 216–220.

[29] S. Vogel, H. Ney, C. Tillmann, HMM-based Word Alignment in Statistical Translation, in: Proceedings of the 16th Conference on Computational Linguistics - Volume 2, COLING '96, 836–841, 1996.

[30] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, R. L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, Comput. Linguist. 19 (2) (1993) 263–311, ISSN 0891-2017.

[31] F. J. Och, H. Ney, A Systematic Comparison of Various Statistical Alignment Models, Comput. Linguist. 29 (1) (2003) 19–51, ISSN 0891-2017.

[32] E. Matusov, R. Zens, H. Ney, Symmetric Word Alignments for Statistical Machine Translation, in: (Ed.), Proceedings of Coling 2004, 219–225, 2004.

[34] N. F. Ayan, B. J. Dorr, A Maximum Entropy Approach to Combining Word Alignments, in: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 96–103, 2006.

[34] N. F. Ayan, B. J. Dorr, A Maximum Entropy Approach to Combining Word Alignments, in: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 96–103, 2006.

[35] K. Ganchev, J. a. V. Gra ca, B. Taskar, Better Alignments = Better Translations?, in: Proceedings of ACL-08: HLT, 986–993, 2008.

[36] T. Baldwin, C. Bannard, T. Tanaka, D. Widdows, An Empirical Model of Multiword Expression Decomposability, in: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, vol. 18, 89–96, 2003.