Automatic Extraction of English-Korean Translations for Constituents of Technical Terms

Jong-Hoon Oh and Key-Sun Choi
Department of Computer Science, Division of EECS, KAIST/KORTERM/BOLA
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea
{rovellia,kschoi}@world.kaist.ac.kr

Abstract*

Technical terms are linguistic realization of a domain concept and their constituents are a component used for representing the concept. Many technical terms are usually multi-word terms and their meaning can be inferred from their constituents. Because a term constituent is usually a morphological unit rather than a conceptual unit in Korean technical terms, we need to first identify conceptual units and then to resolve the proper meaning of the conceptual units in order to properly translate technical terms. For natural language applications to properly handle technical terms, it is necessary to give information about conceptual units and their meaning including homonym, synonym and domain dependency. In this paper, we propose a term constituent alignment algorithm, which extracts such information from bilingual technical term pairs. Our algorithm regards English term constituents as a conceptual unit and then finds its Korean counterpart. Our method shows about 6.1% AER.

1 Introduction

Technical terms are linguistic realization of a domain specific concept and their constituents are a component used for representing the concept (Sager, 1997). Technical terms can be classified into single-word terms, and complex term (or multi-word term) according to the number of their constituents. Single-word terms have one term constituent while complex terms have more than one term constituent. Many Korean technical terms are usually complex terms and their meaning can be inferred from their constituents (Sager, 1997). Therefore it is helpful to identify constituents of technical terms and their meaning in order to understand the meaning of the technical terms and to translate the technical term from one's language to the other. However, a term constituent is usually a morphological unit rather than a conceptual unit[1] in Korean technical terms. Due to the mismatch between a term constituent and a conceptual unit, we need to first identify conceptual units which is a chunk of term constituents representing a domain specific concept ("chunking conceptual units") and then to resolve the proper meaning of the conceptual unit ("resolving meanings") in order to properly understand the meaning of technical terms and to translate them.

In the "chunking conceptual units" stage, it is necessary to determine whether one term constituent represents a concept or not. The decision depends on contexts of term constituents. For example, a Korean technical term, 'seong' can be a conceptual unit by itself when it represents *sex*. But 'seong' in the context of 'hyang-chuk+seong / bun-yeol+jo-jik'[2] (representing *adaxial meristem*) should be recognized as a conceptual unit along with its neighborhood 'hyang-chuk' such as 'hyang-chuk+seong' (*adaxial*). If 'seong' is recognized as a conceptual unit by itself in the context, like 'hyang-chuk (*adaxial*) / seong (*sex*) / bun-yeol+jo-jik (*meris-*

* The first author's current affiliation is with Computational Linguistics Group, National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan

[1] In this paper, a conceptual unit is defined as the linguistic unit representing a domain specific concept.

[2] In this paper, Romanized Korean transcriptions are represented in the quotation mark. In the transcriptions, '+' represents the boundary of term constituents, '-' represents the syllable boundary and '/' represents the boundary of conceptual units.

tem)', we can neither understand the designated meaning of 'hyang-chuk+seong / bun-yeol+jo-jik' ("meristem of a leaf cell in the adaxial area") nor properly translate it.

In the "resolving meanings" stage, *homonym, synonym,* and *domain dependency* of conceptual units should be considered. Sino-Korean affixes are frequently used for coining Korean technical terms and are used as a conceptual unit like single words. Moreover, they are usually *homonym*. For example, a suffix '*-gi*' is used as a term constituent in a biology domain with four senses like *group* (基), *period* (紀), *stage* (期), and *organ* (器). Therefore, disambiguating the sense of such affixes is very important for understanding a Korean technical term.

Many Korean technical terms are from foreign origin. These technical terms become Korean technical terms with various translation ways – 1) translation with pure Korean words, 2) translation with Sino-Korean words, 3) transliteration, 4) combinations of the three ways. Moreover, each translation way produces some variations. For example, *abdominal* is translated into three different Korean terms like '*bok-bu*', '*bok*', and '*bae*', but they indicate the same meaning; in other words, they are synonym. *abdominal* is translated into two Sino-Korean terms like '*bok-bu* (腹部)' and '*bok* (服)', and one pure Korean term, '*bae*'. Capturing synonym, therefore, is important for understanding meaning of technical terms.

Depending on domain of technical terms, translations of conceptual units can be different. For example, the meaning of *cell* in chemistry, physics, and electricity is usually "A single unit that converts radiant energy into electric energy", while that in biology is usually "The smallest structural unit of an organism". In each case, *cell* is differently translated into Korean terms 'jeon-ji' (in chemistry, physics, and electricity domain), and 'se-po' (biology domain).

For natural language applications to properly handle technical terms, it is necessary to give information about conceptual units and their meaning including homonym, synonym and domain dependency. In this paper, we propose a term constituent alignment algorithm, which extracts such information from bilingual technical term pairs. In our algorithm, one or more than one English term constituents are regarded as a conceptual unit. Therefore, the main objective of our algorithm is to recognize conceptual units of Korean technical terms corresponding to an English term constituent in English-Korean translation pairs of technical terms.

The recognized bilingual conceptual units give contextual information, which supports decision whether certain term constituent tends to be used as a conceptual unit by itself or not. Homonym and synonym can be handled by finding the correspondence between English and Korean conceptual units. Because English and Korean conceptual units indicating the same concept will be linked to each other, we can easily find homonym and synonym from the relations. For example, the homonym '*gi*' will be linked to four different English conceptual units. In the same manner, we can capture three relations between the English conceptual unit *abdominal* and its counterparts 'bok-bu', 'bok', and 'bae'. The three Korean counterparts can be clustered as synonyms by means of their corresponding English conceptual unit, like {'bok-bu', 'bok', 'bae'}. Moreover, domain dependency of conceptual units can be handled by the relations because extracted relations for certain English conceptual unit, which has domain dependency, will be different depending on domains.

This paper organized as follows. In section 2, we will describe the related works. Section 3 shows details of our method. Section 4 deals with experiments. Conclusion and future works are drawn in sections 5.

2 Related Works

One of the well-known alignment techniques is the one based on statistical machine translation models. It was initially proposed by (Brown et al., 1993) and, more recently, have been intensively studied by several research groups (Germann et al., 2001; Och et al., 2003). It is used for finding sentence, phrase, and word-level correspondences from parallel texts. It can be formulated as equation (1). For the give source text, *S*, it finds the most probable alignment set, *A*, and target text, *T*.

$$p(T \mid S) = \sum_{a \in A} p(T, a \mid S) \qquad (1)$$

Brown (Brown et al., 1993) proposed five alignment models, called IBM Model, for an English-French alignment task based on equa-

tion (1). Equation (2) describes the IBM Model 1. It is modeled by two assumptions - $P(F|E)$ depends on word translation probability $t(f_j|e_i)$ and one English word was aligned to one French word (1:1 alignment). $t(f_j|e_i)$ is estimated by EM algorithm.

$$p(F \mid E) = C_{l,m} \prod_{j=1}^{m} \sum_{i=1}^{l} t(f_j \mid e_i) \quad (2)$$

where, m represents the length of F, l represents the length of E, and $C_{l,m}$ is a constant value determined by l (the length of E) and m (the length of F).

IBM Model 2 considers distortion (How likely is a source language word in position i to align to a target language word in position j). IBM Model 3 adopts fertility (How likely is a source language word to align to k target language words) as its parameter for *1:n* alignment. IBM Model 4 and 5 make use of relative distortion, word classes and variables to avoid deficiency.

There is another stream of studies on alignment. (Chen et al., 1993; Gale et al., 1993) proposed sentence alignment techniques based on dynamic programming, using sentence length and lexical mapping information. (Haruno et al., 1996; Kay et al., 1993) applied iterative refinement algorithms to sentence level alignment tasks.

In this paper, we propose an alignment algorithm between English and Korean conceptual units (or between English and Korean term constituents) in English-Korean technical term pairs based on IBM Model (Brown et al., 1993). Unlike IBM Model, our alignment model can deal with *n:1* alignment. While the IBM Model aimed to word-level alignment of parallel texts, our method focuses on word- and morphology-level alignment of English-Korean term pairs. Moreover, our algorithm reflects the translation properties of English-to-Korean technical term pairs in a bilingual dictionary.

3 Term Constituent Alignment

For term constituent alignment, we use biology, chemistry and physics dictionaries where term constituents are manually segmented and their part-of-speech is manually assigned. For example, the Korean counterpart of *crop growth rate* is 'jak-mul + seng-jang + yul' and its three term constituents are 'jak-mul', 'seng-jang', and 'yul' where the first two are a noun and the last one is a suffix.

The problem can be defined as finding correspondence between English and Korean term constituents as described in equation (3). For a given English term $E=e_1,\ldots,e_n$, composed of n English term constituents and its corresponding Korean term $K=k_1,\ldots,k_m$, composed of m Korean term constituents, the task is to find alignment set, $A=\{a_1,\ldots,a_t;a_p=(e_{i,i+w(p)}, k_{j(p)})\}$, maximizing probability $P(A|K,E)$, where e_i is the i^{th} term constituent of E, k_j is the j^{th} term constituent of K, and a_p represents the p^{th} alignment relation between English and Korean term constituents. Note that $a_p=(e_{i,i+w(p)}, k_{j(p)})$ $(w \geq 0)$ represents an alignment relation between English term constituents $e_i,\ldots,e_{i+w}$ and Korean term constituent k_j. For example, there are two alignment relations for English term *female sex hormone* and Korean term 'ja-seong + ho-leu-mon', like a_1 =($e_{1,2(1)}$=*female sex*, $k_{1(1)}$='ja-seong') and a_2 =($e_{1(2)}$=*hormone*, $k_{2(2)}$='ho-leu-mon')

$$A^* = \arg\max_{A} P(A \mid K, E) \quad (3)$$

3.1 Statistical Modeling

In this section, first, we describe two translation properties (or constraints), derived from analysis of the alignment tendency between English-Korean term constituents and then describe how to apply these properties to statistical modeling of term constituent alignment.

We randomly sample 20% data of English-Korean term pairs in each technical dictionary and finds two properties "Cross alignment appears in some conditions"[3] and "Null Alignment hardly appears"[4] by analyzing the sampled data.

Constraint 1: Cross alignment is partly allowed.

Let alignment units in a source language be s_i, $s_j(i<j)$, where i and j are the index of the source language, and those in a target language be t_q, t_r $(q<r)$, where q and r are the index in the target language. Then alignment $a_i=(s_i,t_r)$, and $a_j=(s_j,t_q)$ are called cross alignment. Because a sentence structure of Korean is different from

[3] Among analyzed data, 1.3% for biology, 0.1% for physics and 5.65% for chemistry show cross alignment.

[4] Among analyzed data, 0.8% for biology, 0.2% for physics and 0.1% for chemistry show null alignment.

that of English, cross-alignment between English and Korean words frequently occurs in parallel sentences (Shin et al., 1995). For alignment between term constituents, however, most alignment relations are derived from sequential alignment because technical terms, which are usually noun phrases, share the similar structure, say modifier and modifee, in both languages. Sometimes there is cross-alignment because of the preposition in an English term such as *of*. In that case, we allow cross-alignment. For example, there is a cross-alignment relation such as a_1 = (e_2 = *blood*, k_1 = 'hyeol-aek') and a_2 = (e_1 = *clotting*, k_2 = 'eung-go') between the English term *clotting of blood* and its Korean translation 'hyeol-aek + eung-go'. Note that we do not consider the preposition *of* as an alignment unit in that case. English-Korean term pairs representing a name of chemical compounds usually show cross-alignment and 1:1 alignment. To deal with this case, we allow cross-alignment when the number of English term constituents and that of Korean term constituents are same. With the constraint 1, sequential alignment is performed except the above two cases.

Constraint 2: Null Alignment is not allowed.

Constraint 2 means that all English and Korean term constituents should be aligned. Because, term pairs consist of an English term and its translated Korean term, we assume that all constituents should be aligned. Null alignment means that an alignment unit in one side is aligned to nothing in the other side. For example, for *Dutch elm disease* and 'ne-deol-lan-deu (*Dutch*) / neu-leup-na-mu (*elm*) / che-gwan (*sieve tube*) / byeong (*disease*)', there is no English term constituent to be aligned to the Korean term constituent 'che-gwan (*sieve tube*)'. Because, null alignment, however, does not frequently appear in term constituent alignment (only the 0.1%~0.8% data among analyzed data), we do not consider null alignment in our algorithm.

$$P(A \mid K, E) = \prod_{l=1}^{t} p(a_l \mid k_{j(l)}, e_{i,i+w(l)}) \times a(i \mid j, n, m, t) \quad (4)$$

$$p(a_l \mid k_{j(l)}, e_{i,i+w(l)}) = p(a_l \mid k^t{}_{j(l)}, k^w{}_{j(l)}, e_{i,i+w(l)}) \approx p(k^t{}_j \mid e_{i,i+w}) \times p(k^w{}_j \mid k^t{}_j, e_{i,i+w}) \quad (5)$$

By the constraints, equation (3) can be represented as equation (4). In equation (4), n, m, and t represent the number of English term constituents, the number of Korean term constituents and the number of alignment relations between term constituents. In equation (4), $a(i|j,n,t)$ represents position information, which is a binary-valued function and supports the constraint 1. $a(i|j,n,m,t) = 0$ when $a_p = a(e_{i,i+w(p)}, k_{j(p)})$ is cross-alignment, which is not allowed by constraint 1, otherwise $a(i|j,n,m,t) = 1$.

In equation (4), $p(a_l|k_{j(l)},e_{i,i+w(l)})$ are estimated by equation (5). In equation (5), $k_{j(l)}$ is represented by k^w_j and k^t_j where k^w_j and k^t_j are lexical information and part of speech information of the j^{th} Korean term constituent, respectively.

3.2 Parameter Estimation with EM Algorithm

Parameters, $p(k^t_j|e_{i,i+w})$ and $p(k^w_j|\ k^t_j,\ e_{i,i+w})$, in equation (5) are estimated with EM (Expectation-Maximization) algorithm. EM algorithm is the technique for parameter estimation of generic statistical distributions in presence of incomplete data (Dempster et al., 1997). The main goal of EM is to obtain the estimated parameters that give maximum likelihood to the input (incomplete) data. The basic idea underlying the EM algorithm is to iterate through a series of expectation (E-step) and maximization (M-step) steps where the estimation of the parameters of the model is progressively refined until convergence (Lopez et al., 1999).

In this paper, parameters are estimated through two steps, called "initial parameter estimation" and "iterative parameter estimation". In the initial parameter estimation step, the initial parameters are determined by seed data. Seed data, which contains alignment relations derived from $E=e_1,\ldots,e_n$ and E's Korean translation $K=k_1,\ldots,k_m$, where $n=1$ or $m=1$, was selected among data for term constituent alignment. In the condition of $n=1$ or $m=1$, English technical terms or Korean technical terms are a conceptual unit by itself. In other words, alignment relations can be directly extracted from the English-Korean term pairs if

there is only one English term constituent or only one Korean term constituent. With the seed data we can get the initial alignment relation set $A(0)$ and then the initial parameter $\theta(0)$ is estimated with $A(0)$, where $A(k)$ represents the alignment relation set and $\theta(k)$ represents the estimated parameter set derived from the k^{th} iteration. Note that $A=\{a_1,....,a_t;a_p=(e_{i,i+w(p)},k_{j(p)})\}$ and $\theta=\{p(k^t_j|e_{i,i+w}), p(k^w_j|\ k^t_j, e_{i,i+w})\}$.

In the iterative parameter estimation step, $A(k)$ is determined by $\theta(k\text{-}1)$ in E-step and $\theta(k)$ is estimated by $A(k)$ in M-step using the whole data until $\theta(k)$ converges. E-step and M-step can be represented as equation (6)

$$E-step: A(k)=\arg\max_A p(A|E,K;\theta(k-1))$$
$$M-step: \theta(k)=\arg\max_\theta p(\theta|A(k)) \quad (6)$$

$p(k^t_j|e_i)$ and $p(k^w_j|\ k^t_j, e_i)$ are estimated in the k^{th} iteration as equation (7) and (8), respectively. In order to prevent zero probability, the Laplace smoothing method (Manning et al., 1999) is applied to equation (7) and (8).

$$p(k^t{}_j \mid e_{i,i+w}; A(k)) = \frac{1+C(k^t{}_j, e_{i,i+w}; A(k))}{|E|+C(e_{i,i+w}; A(k))} \quad (7)$$

$$p(k^w{}_j \mid k^t{}_j, e_{i,i+w}; A(k)) = \frac{1+C(k^w{}_j, k^t{}_j, e_{i,i+w}; A(k))}{|T|+|E|+C(k^t{}_j, e_{i,i+w}; A(k))} \quad (8)$$

where $C(x)$ represents frequency of x, |E| represents the number of unique English term constituents in $A(k)$, |T| represents the number of unique POS tags of Korean term constituents in $A(k)$.

4 Experiments

For experiments we use three kinds of technical dictionary. They are biology, chemistry, and physics technical dictionaries where Korean term constituents are manually analyzed. The characteristics of experimental data are summarized as Table 1 (Ministry, 2002).

Domain	Seed data	Test data	Total
Biology	8,163	5,668	13,831
Physics	2,757	8,047	10,804
Chemistry	5,353	10,024	15,377

Table 1. Characteristics experimental data (the number of bilingual term pairs)

We compare our model with IBM Model 2 (IBM-2), and IBM Model 4 (IBM-4) implemented by GIZA++ (Och et al., 2003). We evaluate results with the alignment error rate (AER) of Och and Ney (Och et al., 2003), which measures agreement at the level of pairs of term constituents.[5]

$$AER = 1 - \frac{2\times|A\cap G|}{|A|+|G|} \quad (9)$$

where A is the set of term constituent pairs aligned by the automatic system, and G is the set aligned in the gold standard.

4.1 Experimental results

Table 2 shows evaluation results for IBM-2, IBM-4 and our proposed method. In the results precision and AER of our proposed method is higher than those of IBM-4. But recall of our proposed method is lower than that of IBM-4. IBM-4 has strong points in handling cross-alignment and null alignment while our model has strong points in handling *n:1* alignment. The difference between our model and IBM-4 causes the performance gap. Because most alignment type found in the gold standard is *1:1* alignment and *1:n* alignment rather than cross-alignment, null alignment, and *n:1* alignment as described in Table 3, the performance gap between our method and IBM-4 is not so big. IBM-2 shows the worst performance because it can not deal with *1:n* alignment. In other words, IBM-2 does not consider *fertility* as its parameter for estimating the translation probability. Note that *1:n* alignment in the gold standard is about 18%~22% (see Table 3).

Domain	IBM-2	IBM-4	Proposed
Biology	25.0%	7.4%	6.5%
Physics	30.0%	9.6%	5.2%
Chemistry	28.7%	7.6%	6.5%

Table 2. Experimental Results

Type	Biology	Physics	Chem.
Null alignment	0.6%	0.2%	0.2%
Cross alignment	2.1%	0.2%	4.4%
n:1 alignment	2.1%	1.6%	1.2%
1:n alignment	16.5%	21.4%	19.0%
1:1 alignment	78.7%	76.7%	75.3%

Table 3. Alignment types found in the gold standard

When we analyze errors caused by our method, errors are mainly caused by *n:1* alignment and cross-alignment. In order to produce relevant alignment results for *n:1* alignment, we need information indicating that more than one

[5] While (Och et al., 2003) differentiates *sure* and *possible* hand-annotated alignment, our gold-standard comes in only one variety.

English term constituents are used as a conceptual unit. Due to lack of the information, our model has limitation on recovering errors caused by *n:1* alignment. It is necessary to use domain specific corpus as a way of relaxing the problem. Cross alignment, which our model does not allow due to constrain 1, makes errors. Due to the cross alignment, the performance of our method in chemistry and biology is lower than that in physics, where there are few cross alignments in the gold standard.

5 Conclusion

In this paper, we have described an alignment algorithm between English and Korean term constituents. Our alignment algorithm can handle cross alignment, *n:1* alignment and *1:n* alignment between term constituents. Our method shows about 94.7% precision, 93.2% recall and 6.1% alignment error rate. However, there are scopes to improve performance still further. Constraints should be relaxed in order to generalize our model and overcome errors caused by them.

Our method can be applied to handle technical terms in three aspects. First, alignment results produced by our alignment algorithm help a machine translation system to consistently translate new English technical terms to Korean terms by considering domain of the technical terms. Second, alignment results between term constituents can be used for constructing term formation patterns or word formation patterns. Because relations between conceptual units can be extracted from the alignment results, we can construct concept-level term formation patterns using them. Third, the alignment results can be used as a resource for recognizing term variations. Because alignment relations acquired by our alignment model offer information about homonym, synonym and domain dependency, term variations related to certain term constituent can be recognized using them.

Acknowledgement

This work was supported by the Korea Ministry of Science and Technology, the Korea Ministry of Commerce, Industry and Energy, and the Korea Science and Engineering Foundation (KOSEF).

References

Brown P.F., V.S.A. Della Petra, V.J. Della Pietra and R.L. Mercer, “The mathematics of statistical machine translation: parameter estimation”, Computational Linguistics, Vol. 19 No 2, (1993) 263—311

Chen, S, F., Aligning Sentences in Bilingual Corpora Using Lexical information, in proceedings of 31st ACL, (1993) 9—16

Dempster A.P., N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society, 39(1):138, (1977)

Gale, W. A. And Church K.W. A program for aligning sentences in Bilingual Corpora, Computational linguistics, vol 19, no 1, (1993), 75—102

Germann, U. M.Jahr, Knight, K., Marcu, D. And Yamada, K. Fast Decoding and Optimal Decoding for Machine translation, in proceedings of 39th ACL, (2001) 228—235

Haruno M., and Yamazaki, T. High-performance Bilingual Text alignment using Statistical and Dictionary information, in proceedings of 34th ACL, (1996) 131—138

Kay, M. and Roscheisen, M. Text-Translation Alignment, Computational Linguistics, Vol 19, No 1, (1993) 121—142

López de Teruel P. E., José M. García and Manuel E. Acacio. The Parallel EM Algorithm and its Applications in Computer Vision. Parallel and Distributed Processing Techniques and Applications, (1999).

Manning, C.D. and H. Schutze, Foundations of statistical natural language processing, MIT Press (1999)

Ministry of Culture and Tourism, "Forming the foundation of Terminology Standardization”, http://www.korterm.or.kr/, (2002)

Och, Franz Josef and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, Vol 29 (1), (2003), 19—51

Sager, J.C. “Section 1.2.1 Term formation”, in Handbook of terminology management Vol.1, John Benjamins publishing company, (1997)

Shin Jung Ho and Key-Sun Choi (1995), Aligning a parallel Korean-English corpus at word and phrase level, Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium (NLPRS'95), (1995) 223—227