Contextual term equivalent search using domain-driven disambiguation

Caroline Barrière
Centre de Recherche
Informatique de Montréal
Montréal, QC, Canada
barrieca@crim.ca

Pierre André Ménard
Centre de Recherche
Informatique de Montréal
Montréal, QC, Canada
menardpa@crim.ca

Daphnée Azoulay
Laboratoire OLST
Université de Montréal
Montréal, QC, Canada
daphnee.azoulay@umontreal.ca

Abstract

This article presents a domain-driven algorithm for the task of term sense disambiguation (TSD). TSD aims at automatically choosing which term record from a term bank best represents the meaning of a term occurring in a particular context. In a translation environment, finding the contextually appropriate term record is necessary to access the proper equivalent to be used in the target language text. The term bank TERMIUM Plus®, recently published as an open access repository, is chosen as a domain-rich resource for testing our TSD algorithm, using English and French as source and target languages. We devise an experiment using over 1300 English terms found in scientific articles, and show that our domain-driven TSD algorithm is able to bring the best term record, and therefore the best French equivalent, at the average rank of 1.69 compared to a baseline random rank of 3.51.

1 Introduction

We will start this article by introducing, in Section 2, a terminological database called TERMIUM Plus® (referred to as TERMIUM for the rest of the article), which was manually constructed over many years by expert terminologists at the Translation Bureau of Canada. TERMIUM full terminological database has recently been released in an open-data format allowing its use for various research experiments in computational terminology, such as database-wide statistical measures. One particular measure of interest is the notion of similarity between domains, which we present in Section 3.

Section 4 describes our main research contribution, a domain-driven term sense disambiguation (TSD) algorithm. TSD aims at automatically determining which term record from a term bank best represents the meaning of a term given its context. This is a task that translators must perform on a regular basis when translating specialized texts containing specialized terminology. Finding the contextually appropriate term record leads the translator to the proper equivalent to use in his or her translation. Making the algorithm *domain-driven* means that the information that will be used to perform the disambiguation task is the domain information provided in each term record of the term bank. For example, according to TERMIUM, the French equivalent *promontoire* would be proper for the word *head* found in a text segment about the TOPONYMY domain, but the equivalent *tête* would be more appropriate in other domains such as STRING INSTRUMENT, or GOLF.

In Section 5, we present an experiment to evaluate the performances of our algorithm. We will describe the dataset composed of 1500 terms found in abstracts of scientific publications, the human annotation performed to build a gold standard, the algorithm parameter optimisation using a subset of 200 terms, and the final results on the remaining 1300 terms.

Domain-driven TSD is definitely an underexplored task within the Natural Language Processing literature, and Section 6 will give some pointers to only a few related works which rather use domain information as complementing other information for disambiguation. Term disambiguation in general, domain-driven or not, is rarely explored perhaps due to a misconception that terms are monosemous and that disambiguation is not necessary in specialized domains. Although that statement is true of most

Proceedings of the 5th International Workshop on Computational Terminology,
pages 21–29, Osaka, Japan, December 12 2016.

multi-word terms, it is certainly not true of the many single-word terms found in specialized texts which tend to lead to multiple term records.

Finally, in Section 7, we conclude and give an outlook to future work.

2 TERMIUM as an open-data terminological resource

TERMIUM has been in constant expansion for over 20 years, and is the result of much labour from terminologists at the Translation Bureau of Canada. But only since 2014 has TERMIUM been openly and freely available as part of Canada's Open Government initiative[1].

The 2015 open-data version of TERMIUM, used in the current research, contains 1,348,065 records, organized within 2203 domains. Each record corresponds to a particular concept within particular domains, with its multilingual term equivalents. For example, a record for the concept defined as *Irregularity or loss of rhythm, especially of the heartbeat* within the MEDICAL domain, would provide three term equivalents: *arrhythmia* (English), *arythmie* (French), and *arritmia* (Spanish). Even though some terms are quite specific to a single domain, such as *arrhythmia*, some other terms, such as *head*, do belong to 55 domains, including TOPONYMY, SHIP AND BOAT PARTS, STRING INSTRUMENT, METAL FORMING, GOLF, and STOCK EXCHANGE.

Table 1 shows a few single-word terms, in general more polysemous than multi-word terms, chosen to illustrate the variability in the number of records associated to each term (column 2), the number and variety of possible French equivalents (column 3 and 4), and the number and variety of possible domains (column 5 and 6). There is no one-to-one relation between term equivalents and domains. For example the term *resistance* leads to 14 domains and only 2 French equivalents, whereas *quenching* leads to 7 domains and 5 equivalents. Although TERMIUM covers three languages (English, French, Spanish), we will focus on the English/French language pair in this article.

Table 1: Examples of terms in TERMIUM

English term	Nb records	Nb Equiv	Examples of French equivalents	Nb domains	Examples of domains
resistance	14	2	résistance, défense	14	CROP PROTECTION, TEXTILE INDUSTRIES, HORSE HUSBANDRY, PADDLE SPORTS
nucleus	10	3	noyau, nucléus, germe	19	CYTOLOGY, METALS MINING, BEEKEEPING, ARCHEOLOGY, HAND TOOLS
quenching	5	5	surfusion, refroidissement rapide	7	BIOTECHNOLOGY, ENERGY (PHYSICS), GEOPHYSICS, PLASTIC MATERIALS
evolution	3	1	évolution	4	GENETICS, PALEONTOLOGY, MATHEMATICS

When terminologists create term records, they are required to specify domain information. They are also encouraged to include definitions, contexts of usage, and other observations, but it is not mandatory. We calculated simple statistics for English and French, and found that only 14.2 % of the records contained definitions in English and 14.6 % contained definitions in French. These statistics encourage the development of an algorithm which solely use the domain information and does not rely on the definitional information which would only cover a small percentage of the records. Furthermore, restricting the algorithm to domain information makes the algorithm highly portable to other term banks, like EuroTermBank[2] or IATE [3], that would also be structured using records and domains, as typical term banks are.

3 Measuring domain similarity

In our algorithm of domain-driven disambiguation, presented in the next section, it will be important to assess the similarity between domains. For example, if the algorithm tries to disambiguate a term found in the context of GEOLOGY, TERMIUM might offer only two term records, one within the domain of EARTH SCIENCE and the other one within the domain of ANIMAL BEHAVIOUR. In such case, the algorithm

[1] TERMIUM Plus®, Government of Canada, `http://open.canada.ca/data/en/dataset/`.

[2] EuroTermBank can be found at `http://www.eurotermbank.com`

[3] InterActive Terminology for Europe (IATE) can be found at `http://http://iate.europa.eu`

needs some measure of domain similarity to decide between the term records, since neither one refers to the exact same domain as the text.

TERMIUM does provide a simple domain hierarchy with coarse-grained and fine-grained domains, intrinsically showing some similarity between domains. For example, the domain AGRICULTURE would include sub-domains such as CROP PROTECTION, CULTURE OF FRUIT TREES, and GRAIN GROWING, whereas the domain HEALTH AND MEDICINE would include sub-domains such as RESPIRATORY TRACT, ACUPUNCTURE and RADIOTHERAPY. Unfortunately, such hierarchy is not sufficient to measure similarities among the sister domains (e.g., RESPIRATORY TRACT, ACUPUNCTURE) which will be required for our algorithm.

We rather opt for similarity measures commonly used for measuring word collocation strengths, such as Overlap or Point-Wise Mutual Information (PMI), which we will adapt to measure domain similarity. Most measures of collocation strength between two words, W_1 and W_2, rely on three counts: the number of segments (e.g., documents, sentences or fixed-sized text windows) in which W_1 and W_2 occur together, the number of segments in which W_1 occurs, and the number of segments in which W_2 occurs.

We transpose this idea to the terminological database, considering each term record as a possible segment. The similarity between two domains, D_1 and D_2, then refers to their collocation strength, meaning how likely they are to co-occur on a term record.

For example, the domain of GENETICS is present on 4664 records, of which 203 are also assigned to BIOCHEMISTRY. On the other hand, the same domain GENETICS has zero record in common with the domain of REPROGRAPHY. Using these counts, the domain similarity measures will be able to express that GENETICS is more similar to BIOCHEMISTRY than it is to REPROGRAPHY.

The two measures we have tested to compare two domains, D_1 and D_2, are provided below, with $NbRecords$ representing a number of term records.

$$PMI(D_1, D_2) = \frac{NbRecords(D_1, D2)}{NbRecords(D_1) * NbRecords(D_2)} \tag{1}$$

$$OVERLAP(D_1, D_2) = \frac{NbRecords(D_1, D2)}{MIN(NbRecords(D_1), NbRecords(D_2))} \tag{2}$$

In Table 2, we see the results with both the PMI measure (Equation 1) and the Overlap measure (Equation 2) as to the top 10 closest domains to REPROGRAPHY and CYCLING. The lists are slightly different (domains in common between the two measures are highlighted in bold), but it is very hard to provide a real evaluation of these lists until they are actually used in different tasks requiring them. In general, intrinsic evaluation of similarity measures is quite difficult out of context, leading to much subjectivity and therefore low inter-annotator agreement. As our goal is term disambiguation, we will instead perform an extrinsic evaluation, by determining which similarity is best for our task, as we describe in Section 5.3.

Table 2: Examples of closest domains (PMI and Overlap)

Domain	Measure	Closest domains
Reprography	PMI	**NON-IMPACT PRINTING / INTAGLIO PRINTING / POWER TRANSMISSION TECHNIQUES / LITHOGRAPHY, OFFSET PRINTING AND COLLOTYPE / PHOTOGRAPHY / PRINTING PROCESSES - VARIOUS** / INKS AND COLOUR REPRODUCTION (GRAPHIC ARTS) / **OFFICE EQUIPMENT AND SUPPLIES** / BIOMETRICS / OFFICE MACHINERY
	Overlap	**NON-IMPACT PRINTING / PHOTOGRAPHY / OFFICE EQUIPMENT AND SUPPLIES / INTAGLIO PRINTING** / AUDIOVISUAL TECHNIQUES AND EQUIPMENT / **POWER TRANSMISSION TECHNIQUES / LITHOGRAPHY, OFFSET PRINTING AND COLLOTYPE / PRINTING PROCESSES - VARIOUS** / GRAPHIC ARTS AND PRINTING / OFFICE AUTOMATION
Cycling	PMI	**MOTORCYCLES AND SNOWMOBILES / MINING TOPOGRAPHY / MOTORIZED SPORTS / SPORTS EQUIPMENT AND ACCESSORIES / CONSTRUCTION WORKS (RAILROADS) / SHELTERS (HORTICULTURE) / ROADS / TRACK AND FIELD / WHEELS AND TIRES (MOTOR VEHICLES AND BICYCLES)** / SPORTS FACILITIES AND VENUES
	Overlap	**MOTORCYCLES AND SNOWMOBILES / MINING TOPOGRAPHY / MOTORIZED SPORTS / SPORTS EQUIPMENT AND ACCESSORIES / CONSTRUCTION WORKS (RAILROADS) / SHELTERS (HORTICULTURE) / TRACK AND FIELD** / HORSE RACING AND EQUESTRIAN SPORTS / **ROADS / WHEELS AND TIRES (MOTOR VEHICLES AND BICYCLES)**

4 Domain-driven disambiguation algorithm

Assume a textual context C, such as the small paragraph below, and a term T, such as *virus* or *nucleus*, to be disambiguated in order to find its proper French equivalent.

> *Transforming infection of Go/G1-arrested primary mouse kidney cell cultures with simian virus 40 (SV40) induces cells to re-enter the S-phase of the cell cycle. In Go-arrested cells, no p53 is detected, whereas in cells induced to proliferate by infection, a gradual accumulation of p53 complexed to SV40 large T-antigen is observed in the nucleus. Heat treatment of actively proliferating SV40-infected cells leads to inhibition of DNA synthesis and growth arrest. To determine the fate of p53 after heat treatment, proliferating infected cells were exposed to mild heat (42.5C) for increasing lengths of time.*

We present a domain-driven disambiguation algorithm which aims at disambiguating T given context C. There are three important steps to this algorithm which we present in details.

4.1 Extracting profiling terms

Profiling terms are terms found in context C which are representative of its content. For our particular purpose, these profiling terms must be present in TERMIUM as they will serve to further determine the domains conveyed in the text.

The context C is pre-processed through tokenization, lemmatization and POS-tagging[4]. Once the text is lemmatized, we choose the longest sequences of lemmas found as terms in TERMIUM leaving out overlapping shorter terms. For example, the segment *primary mouse kidney cell cultures with simian virus 40 (SV40) induces cells* contains two multi-word TERMIUM terms, *kidney cell culture* and *simian virus 40*, as well as four single-word terms, *primary*, *mouse*, *induce* and *cell*.

The initial set of profiling terms can then be reduced through syntactic and semantic filtering. Section 5.3 will measure the impact of such filtering on the disambiguation task. The syntactic filtering makes use of the POS tagging, allowing to restrict the list of terms to only verbs and nouns (removing the adjective *primary* in the example above), or even to only nouns (further removing the verb *induce* in the example above).

The semantic filtering is based on the degree of polysemy allowed for the profiling terms. The first line of Table 3 shows that 13 profiling terms would be kept if the maximum polysemy allowed was of 10 term records, following a syntactic filter for keeping nouns only. The following lines of Table 3 show how the number of profiling terms reduces significantly as the semantic filter further limits the degree of polysemy. Only two terms are left, *cell cycle* and *kidney cell culture*, when restricting to monosemous terms only. The hypothesis to be later confirmed is that perhaps profiling a text using only its monosemous terms, or slightly polysemous terms, would lead to a better disambiguation overall.

Table 3: Impact on profiling terms when filtering with a threshold on polysemy

Max polysemy	Profiling terms retained
10 records	synthesis, heat, infection, fate, arrest, cell cycle, length, nucleus, inhibition, heat treatment, virus, mouse, kidney cell culture
5 records	synthesis, infection, fate, cell cycle, virus, mouse, kidney cell culture
3 records	synthesis, infection, fate, cell cycle, virus, kidney cell culture
2 records	cell cycle, virus, kidney cell culture
1 record	cell cycle, kidney cell culture

4.2 Building a domain profile

Once a set of profiling terms has been extracted, we can automatically search in TERMIUM for their associated domains. For example, if we take the subset of profiling terms having a maximum number of three term records (see Table 3), we can see their associated domains in Table 4. This information is used to build the actual domain profile, a subset of which is shown in Table 5.

The weight of each domain within the domain profile is based on a simple $tf * idf$ style of weighting, where tf is the number of times a profile term occurs in context C, and where idf is calculated as $\frac{1}{N}$

[4]Stanford Core NLP tagger was used, available at `http://nlp.stanford.edu/software/tagger.shtml`.

Table 4: Domains found on the term records of the profiling terms

Term	Nb Records	Domains
synthesis	3	[BIOTECHNOLOGY, BIOLOGICAL SCIENCES, ARTIFICIAL INTELLIGENCE]
infection	3	[HUMAN DISEASES, EPIDEMIOLOGY, BREWING AND MALTING, IT SECURITY]
fate	3	[AGRICULTURAL CHEMICALS, MENTAL DISORDERS, BANKING, ENVIRONMENTAL STUDIES AND ANALYSES]
cell cycle	1	[BIOTECHNOLOGY, CYTOLOGY]
virus	2	[MICROBIOLOGY AND PARASITOLOGY, COMPUTER PROGRAMS AND PROGRAMMING, IT SECURITY]
kidney cell culture	1	[CYTOLOGY]

where N is the number of domains associated with the profile term within TERMIUM. For example, the idf for *fate* is 0.25 since it occurs in four domains. A domain's total weight (column 3) is the sum of the profile term weights contributing to it. The contributing terms to each domain are shown in column 4.

Table 5: Domain Profile

Domain Profile (DP_i)	Domain name	weight (W_{DP_i})	Contributing term
DP_1	CYTOLOGY	1.5	kidney cell culture (1.0), cell cycle (0.5)
DP_2	BIOTECHNOLOGY	0.83	synthesis (0.33), cell cycle (0.5)
DP_3	IT SECURITY	0.58	infection (0.33), virus (0.25)
DP_4	ARTIFICIAL INTELLIGENCE	0.33	synthesis (0.33)
DP_5	BIOLOGICAL SCIENCES	0.33	synthesis (0.33)
...	...	...	...
DP_{13}	HUMAN DISEASES	0.25	infection (0.25)
DP_{14}	MENTAL DISORDERS	0.25	fate (0.25)

As we previously discussed in Section 4.1, both syntactic and semantic filters will affect the set of profile terms, which consequently will affect the domain profile. Table 6 shows different domain profiles associated with different combinations of syntactic and semantic filters on the profiling terms. It is quite difficult and somewhat subjective to assess the domain profiles directly, and the impact of the various parameters will rather be measured on the disambiguation task.

Table 6: Examples of corresponding domain profiles

Syntactic Filter	Semantic Filter	Top 5 Domains
None	Max 10 records	GENERAL VOCABULARY (2.78) CYTOLOGY (1.55), TRANSLATION (GENERAL) (1.07), BIOTECHNOLOGY (1.05), DENTISTRY (1.0)
Nouns	Max 20 records	CYTOLOGY (1.55), BIOTECHNOLOGY (0.89), IT SECURITY (0.65), BIOLOGICAL SCIENCES (0.42), COMPUTER PROGRAMS AND PROGRAMMING (0.39)
Nouns	Max 3 records	CYTOLOGY (1.5), BIOTECHNOLOGY (0.83), IT SECURITY (0.58), ARTIFICIAL INTELLIGENCE (0.33), BIOLOGICAL SCIENCES (0.33)
Nouns-Verbs	Max 1 records	CYTOLOGY (1.5), DENTISTRY (1.0), BIOTECHNOLOGY (0.5)

4.3 Establishing the most likely domain for term T

The last step in our algorithm requires a domain-to-domain similarity matrix, M, providing a similarity measure for each domain pair found in TERMIUM. Such matrix M will show, for example, that BIOLOGY is similar to ZOOLOGY, but unrelated to FINANCIAL MARKET. Section 3 discussed how to measure domain similarity.

Having pre-calculated M for all domain pairs in TERMIUM, we use M to establish the most likely domain for T. Let's refer to a possible domain of T as D_i, among N possible domains $D_1..D_N$. For each D_i, we calculate its domain strength by summing its similarity to each of the X domains $DP_1..DP_X$ making up the domain profile of context C. Each similarity, $M(D_i, DP_j)$, is further weighted by the score of each domain in the profile (see for example column 3 of Table 5). Equation 3 shows the calculation.

$$DomainStrength(D_i) = \sum_{j=1}^{X} M(D_i, DP_j) * W_{DP_j} \quad (3)$$

As an example, we show in Table 7 the top 5 term records obtained for the term *nucleus*, after performing the calculation above for each of its possible domains. Note that a term record is often associated to more than one domain. In such cases, the score of the term record is set to the average domain strength of its domains.

If the algorithm performs well, the highest score (rank 1) should be the correct term record. The experiment described in the next section presents a proper evaluation of the performances of the algorithm.

Table 7: Examples of ranking term records for the term *nucleus*

Record Rank	Domains	Score	French
1	[BIOTECHNOLOGY, CYTOLOGY]	1.378	noyau
2	[MOLECULAR BIOLOGY, ATOMIC PHYSICS]	0.118	noyau
3	[METALLOGRAPHY]	0.036	germe
4	[COMPUTER PROGRAMS AND PROGRAMMING]	0.032	noyau
5	[AQUACULTURE, MARINE BIOLOGY, JEWELLERY, MOLLUSKS, ECHINODERMS AND PROCHORDATES]	0.008	nucleus

5 Experiment: term disambiguation from scientific abstracts

This section describes a term sense disambiguation experiment applied on a large dataset of 1500 terms. We first describe the dataset and the human domain annotation performed to obtain our gold standard. Then, we describe the various parameter adjustments performed on a subset of 200 annotated terms which we used as our development set. Finally, we describe the results of the fine-tuned algorithm on the remaining set of 1300 terms.

5.1 Dataset

For our experiment, we use a dataset described in (Carpuat et al., 2012) of scientific abstracts from various journals published by the Research Press of National Research Council of Canada[5]. In total, the dataset contains 3347 abstracts from eleven journals covering topics of biology, earth science, chemistry, and more.

The short example text used in Section 4 for describing the algorithm was taken from an abstract of an article in a biology journal. The abstracts are usually followed by three to five author-provided specialized keywords (terms). We earlier discussed how *nucleus* could be such a possible term from this abstract.

For a term to be included in our dataset, we require that it be present in TERMIUM and be polysemous. Given these constraints, we gathered 1500 terms for testing from which 200 terms were used for development. The degree of polysemy varies largely among the dataset. To provide a few statistics, we measured that 38.3% of the terms led to only 2 records, 22.8% led to 3 and 4 records, but there are also 20.7% of the terms leading to more than 10 records, providing quite a challenge for automatic disambiguation.

5.2 Annotation effort

The human effort required to choose the proper term record corresponds to what a translator needs to do when searching for the proper equivalent in a term bank to best convey the meaning of a term occurring in a particular context. To simulate this effort and provide a gold standard for our task, a master's student in terminology annotated 1500 terms chosen randomly from the various scientific abstracts. For each term, the annotator had to indicate the most contextually appropriate TERMIUM term record. Disambiguation of the 1500 terms represented a 40 hour effort. The annotator reported that for the majority of terms, finding the appropriate record was done easily by using the domain information.

Through the annotation, some interesting cases came about. First, the annotator noticed duplications within TERMIUM as some records corresponded to similar domains, had similar definitions, and usually

[5]The list of journals can be seen at `http://nrcresearchpress.com/`, and the dataset can be found at `http://www.umiacs.umd.edu/~hal/damt/` as part of JHU Summer Workshop in Domain Adaptation in Machine Translation.

also shared the same French equivalent, indicating that they could have been grouped into a single record. In such cases, both records were indicated as correct in the gold standard.

Second, the annotator found cases when no term record contained a domain which exactly matched the context of occurrence of the term. The followed annotation guideline was that if a term record contained a domain that was somewhat related to the expected domain and that it also contained an appropriate French equivalent, then it should be chosen as the correct term record. For example, if the term *sugar* was expected in the BIOCHEMISTRY domain, but TERMIUM contained FOOD INDUSTRY as the closest domain on a term record, that term record was chosen as correct. This guideline made sense for our particular gold standard since the domain-driven algorithm used in our experiment only has access to the domains provided in TERMIUM and chooses the best one among them.

Finally, the annotator found cases when no term record contained domains close enough to the context, nor did any term record contained any contextually suitable definition or proper equivalent. For example, if the term *hedgehog* was expected in the GENETICS domain, but TERMIUM only contained term records within the NAVAL DOCKYARDS and FIELD ENGINEERING domains, neither seemed close enough to be appropriate. The guideline for such cases was to annotate the term as "no record". The terms corresponding to this annotation were discarded from the current experiment, although they could be used in future work toward the evaluation of an extended algorithm further able to determine when no record is appropriate.

A more extensive annotation effort would involve multiple annotators and a measure of inter-annotator agreement. For the current experiment, given resource limitations, a single annotator was asked to perform the task. Although limited, it was deemed nonetheless appropriate as we performed an intra-annotator test, with the same annotator waiting two weeks between two annotation efforts of a same sample of 50 terms. Such test showed that the annotator's decisions were reliable enough to be used as a gold standard.

5.3 Parameter tuning

The original development dataset of 200 terms had to be reduced to 178 terms for two reasons. The first one is that some of the terms led to the annotation "no record" by the annotator (as mentioned in the previous section). The second one is that we found out too late that the on-line version of TERMIUM used for the human annotation was not exactly the same as the open-data one used for the experimentation, leading to some choices of domains made by the annotator which were actually not among the ones present in the open-data version.

We have seen in Section 4, as we described the three steps of the algorithm, that multiple parameters could influence the final results. First, for the domain similarity matrix (Section 3), we tested two possible similarity measures, PMI and Overlap. Second, for the profiling terms (Section 4.1), we suggested imposing syntactic restrictions, keeping only words tagged with specific parts of speech, as well as imposing semantic restrictions, keeping only terms leading to a maximum number of records within TERMIUM. As for syntactic restrictions, we tried without filter, with noun and verb filter and with noun-only filter. As for the semantic restrictions, we tried with maximum number of records of 1, 2, 5 and 10 records.

Our development set is sufficiently small, and our disambiguation process sufficiently fast, that we could vary all the different parameters in combination. Such experiment showed that all parameters were influential in the quality of the results: (1) the domain similarity measure (PMI largely outperforming Overlap), (2) the degree of polysemy of the profile terms (lower, even complete monosemy, was best) and (3) the syntactic constraint put on the profile terms (keeping only nouns outperformed the other options). The variation in results was quite significant. For 178 terms, with the best combination of parameters (PMI, monosemic terms, keeping only nouns), we have an average rank of 1.62, and with the worst combination (Overlap, polysemy up to 10 records, keeping nouns and verbs) we have an average rank of 2.30. The rank of a random baseline would be of 3.84, given the average number of records being 7.68. The best combination of parameters will be used for the evaluation in the next section.

5.4 Evaluation

In the same way that our development set was reduced from 200 terms to 178 terms given the various cases encountered during the annotation effort, the evaluation dataset also ended up being reduced from 1300 terms to 1175 terms.

On these 1175 terms, the random rank was evaluated at 3.51 and the domain-driven disambiguation algorithm, using the best combination of parameters (PMI, monosemic terms, keeping only nouns), reduced that rank to 1.69, showing a significant improvement.

In Table 8, we further show the proportion of terms leading to the various ranks. It is interesting to see that for almost 75% of the terms, the algorithm succeeds in bringing the best record to the top (rank 1).

Table 8: Percentage of terms per rank obtained for the correct answer

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	> Rank 5
74.5 %	13.8 %	4.0 %	2.0 %	1.9 %	3.2 %

From an application point of view, the most interesting result lies in the disambiguation capability of the algorithm for largely polysemous terms, since those would be time-consuming for translators, requiring them to go through multiple records to find the appropriate record given the context. For example, the term *disturbance* has 12 term records, *roughness* has 13, *binding* has 25 and *cluster* has 40, and for all those terms, the algorithm was able to bring the contextually appropriate term record to the top rank.

6 Related work

Given that a term bank does not always contain definitions of terms, we have restricted our algorithm to the use of domain information, and opted for a domain-driven disambiguation approach. This is quite different from many unsupervised word sense disambiguation approaches, which make use of the definitions of the senses for comparing them with the context of use of the ambiguous word. Such definition-based approach is often called Lesk-like, given its root in (Lesk, 1986), and later modified into multiple variations (Vasilescu et al., 2004).

Some work has focused on subject fields in Wordnet. Integration of subject field codes (Magnini and Cavaglia, 2000) in WordNet, also called WordNet domains[6], has led to some domain-driven algorithms for word sense disambiguation (Magnini et al., 2002). In (Gliozzo et al., 2004), a domain relevance estimation is performed to assign domains to text. Their domain relevance estimation is similar in intent to our context profiling, but performed with a supervised machine learning approach.

The present research builds on our previous work (Barrière, 2010) which developed a domain-driven disambiguation algorithm using the Grand Dictionnaire Terminologique (GDT)[7] as term bank. Unfortunately, our previous results were unreproducible for other researchers in the community since the GDT is not published in an open-data format. The recent release of TERMIUM in an open-data format allowed us to implement, test, and further refine our earlier algorithm. Our evaluation in (Barrière, 2010) was also not too convincing, since we measured success based on the finding of the proper equivalent and not the proper term record. Given that multiple records could lead to the same term equivalent, such evaluation was quite optimistic. Our current annotation effort at the record level allows us to provide a more realistic evaluation. Yet, given the refinements we introduced in the current algorithm, our realistic results at 1.69 average rank is better than the earlier optimistic result at 2.0 average rank.

We are not aware of other work addressing the term sense disambiguation problem as such, using domain-driven methods to counter the lacking presence of definitions in term banks.

[6]WordNet domains are available at `http://wndomains.fbk.eu/`.

[7]The Grand Dictionnaire Terminologique is published by the Office québécois de la langue française and can be consulted online at `http://gdt.oqlf.gouv.qc.ca/`.

7 Conclusion and future work

We presented TERMIUM Plus®, a resource from the Translation Bureau of Canada. TERMIUM is intended for translators as end users, but its recently released open-data version could make it a resource of much interest to the computational terminology research community. We showed its usefulness in a term equivalent search experiment. We presented a domain-driven disambiguation algorithm, relying on domain similarity estimations on the overall resource, and on context profiling. Our algorithm significantly reduced the average rank of the appropriate equivalent from 3.51 (baseline of random assignment) to 1.69, on an unseen dataset of 1175 terms.

From an application point of view, our domain-driven term sense disambiguation algorithm could be used for automatic pre-translation of specialized terms in text. Or perhaps, the algorithm could point out to possible translation errors, in cases of a discrepancy between the translator's choice and what seems to be the best equivalent according to automatic disambiguation. To support this idea, let us point out that we noticed a few examples in our dataset where the actual term equivalent found on the appropriate record was not the one chosen by the translator.

Even though TERMIUM was the chosen term bank for our experiment, it would be very interesting to put our domain-driven disambiguation algorithm to the test using other term banks, such as EuroTermBank or IATE, which would be structured given their own set of domains. Wikipedia categories, although much more loosely defined than the domains in TERMIUM, and in a collaborative manner rather than a curated one, are also worth investigating for implementing a category-driven disambiguation approach. We can also explore the approach described by (Pang and Biuk-Aghai, 2010) to determine category similarity. More recently, (Gella et al., 2014) have suggested a method for mapping WordNet domains with Wikipedia Categories, which would perhaps allow us to explore the combination of both resources within a text profiling and domain-driven disambiguation approach.

References

Caroline Barrière. 2010. Recherche contextuelle d'équivalents en banque de terminologie. *Traitement Automatique des Langues Naturelles (TALN'2010)*.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Ales Tamchyna, Katharine Henry, and Rachel Rudinger. 2012. Domain Adaptation in Machine Translation: Final Report. In *2012 Johns Hopkins Summer Workshop Final Report*.

Spandana Gella, Carlo Strapparava, and Vivi Nastase. 2014. Mapping WordNet Domains , WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources. pages 1117–1121.

Alfio Gliozzo, Bernardo Magnini, and Carlo Strapparava. 2004. Unsupervised Domain Relevance Estimation for Word Sense Disambiguation. *Proc. of the 2004 Conference on EMNLP*, pages 380–387.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *SIGDOC'86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA.

Bernardo Magnini and G Cavaglia. 2000. Integrating Subject Field Codes into WordNet. *In Proceedings of LREC 2000*, pages 1413–1418.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4):359–373.

Cheong-Iao Pang and Robert P Biuk-Aghai. 2010. A Method for Category Similarity Calculation in Wikis. In *WikiSym'10*, Gdansk, Poland.

Florentina Vasilescu, Philippe Langlais, and Guy Lapalme. 2004. Evaluating Variants of the Lesk Approach for Disambiguating Words. In *Language Resources and Evaluation (LREC)*, pages 633–636.