Identifying paraphrases between technical and lay corpora

Louise Deléger, Pierre Zweigenbaum

LIMSI-CNRS
BP133, F-91403 Orsay, France
louise.deleger@limsi.fr, pz@limsi.fr

Abstract

In previous work, we presented a preliminary study to identify paraphrases between technical and lay discourse types from medical corpora dedicated to the French language. In this paper, we test the hypothesis that the same kinds of paraphrases as for French can be detected between English technical and lay discourse types and report the adaptation of our method from French to English. Starting from the constitution of monolingual comparable corpora, we extract two kinds of paraphrases: paraphrases between nominalizations and verbal constructions and paraphrases between neo-classical compounds and modern-language phrases. We do this relying on morphological resources and a set of extraction rules we adapt from the original approach for French. Results show that paraphrases could be identified with a rather good precision, and that these types of paraphrase are relevant in the context of the opposition between technical and lay discourse types. These observations are consistent with the results obtained for French, which demonstrates the portability of the approach as well as the similarity of the two languages as regards the use of those kinds of expressions in technical and lay discourse types.

1. Introduction

Paraphrases can be a useful resource for many natural language processing applications, including information retrieval (Ibrahim et al., 2003), information extraction (Shinyama and Sekine, 2003), text simplification (Elhadad and Sutaria, 2007) and authoring aids (Max, 2008). Most existing approaches (Barzilay and Lee, 2003; Shinyama and Sekine, 2003; Pasca and Dienes, 2005) aim at extracting undifferentiated paraphrases, regardless of the kind of discourse they belong to. A more fine-grained characterization of paraphrases would give insight into their context of use. In this regard, distinguishing between technical and lay discourse types would be especially helpful for text simplification or authoring aid, as it would allow to choose one paraphrase over another according to the target audience. This is the goal of the present work.

Very few studies have looked for paraphrases between different discourse types. In a medical context, Elhadad and Sutaria Elhadad and Sutaria (2007) extracted lay paraphrases of technical terms from a comparable corpus of medical abstracts. We presented a preliminary study to identify paraphrases between technical and lay discourse types, also from medical corpora, and dedicated to the French language (Deléger and Zweigenbaum, 2009). In this paper, we test the hypothesis that the same kinds of paraphrases as for French can be detected between English technical and lay discourse types and report the adaptation of our method from French to English, including variations to improve the original approach.

Several approaches are possible when dealing with paraphrase extraction from corpora. We classify them according to the type of corpora they rely on. Methods can use plain corpora, such as (Jacquemin, 1999) who detects term variants or (Pasca and Dienes, 2005) who extract paraphrases from random Web documents. Some approaches rely on parallel corpora (*i.e.* different translations or versions of the same texts) which can be either monolingual (Barzilay and McKeown, 2001) or bilingual (Bannard and Callison-Burch, 2005; Max, 2008). Comparable corpora are another useful source of paraphrases (Barzilay and Lee, 2003; Shinyama and Sekine, 2003; Elhadad and Sutaria, 2007). In this regard, only closely related corpora have been used, especially and almost exclusively corpora of news sources reporting the same events (Barzilay and Lee, 2003; Shinyama and Sekine, 2003).

Here, the nature of the paraphrases we are interested in (i.e. paraphrases between two types of discourse) calls for corpora with specific properties contrasting a set of documents (here, documents written in a technical discourse) to another (here, documents written in a lay discourse). This set of documents can be either parallel corpora or comparable corpora. As parallel corpora are very scarce resources, especially in our domain of application (the medical field) and with documents from two different discourse types, the natural choice was to use comparable corpora.

We describe our experiment, starting from the constitution of monolingual comparable corpora (section 2), then detailing the extraction of technical vs. lay paraphrases (section 3). We expose our results (section 4) and discuss them (section 5).

2. Building comparable corpora

Following our methodology in (Deléger and Zweigenbaum, 2009), we built two monolingual comparable corpora of English-language medical specialized and lay texts from the Web, in the domains of cancer and diabetes. The cancer corpus was directly compiled from a website providing comparable texts: the National Institute for Health and Clinical Excellence website[1] which publishes guidelines for physicians and their lay versions for the general public. We only selected documents dealing with the cancer topic. This corpus is parallel at the document level (but not at the sentence or word level). Since such ready-made comparable corpora are rare resources, we also gathered a second corpus from various sources through a guided search:

[1] http://www.nice.org.uk/

Original French rule	Adapted English rule
(DET) N_1 PREP (DET) $N_2 \rightarrow V_1$ (DET) N_2	(DET) N_2 $N_1 \rightarrow V_1$ (DET) N_2
(DET) N_1 PREP (DET) N_2 $A_3 \rightarrow V_1$ (DET) N_2 A_3	(DET) A_3 N_2 $N_1 \rightarrow V_1$ (DET) A_3 N_2
(DET) N_1 $A_2 \rightarrow$ (DET) N_2 V_1	(DET) A_2 $N_1 \rightarrow$ (DET) N_2 V_1

Table 2: Rule adaptation (a shared index indicates equality or synonymy. N=noun, V=verb, A=adjective, PREP=preposition, DET=determiner, 1 in index = pair of deverbal noun and verb)

Original French rule	Adapted English rule
C $\rightarrow$ (N (A) (PREP)) C_1 W$^{0-4}$ C_2	C $\rightarrow$ ((A) N (PREP)) C_1 W$^{0-4}$ C_2

Table 3: Rule adaptation (C is a neo-classical compound, C_1 and C_2 are the modern-language equivalents of the components of C, N is a noun, PREP a preposition, A an adjective and W an arbitrary word)

we thus identified and queried a number of relevant websites, including the search engine for biomedical articles PubMed[2], the online medical manual Merck[3], and clinical guidelines from the National Guideline Clearinghouse[4] for the technical side of the corpus; and websites dedicated to the general public (MedlinePlus[5], WebMD[6], NetDoctor[7]) for the lay side of the corpus. Table 1 gives the sizes of the corpora, in terms of documents and words for each side of the corpora.

	Diabetes		**Cancer**	
	techn.	*lay*	*techn.*	*lay*
Documents	41	1,512	25	25
Words	360,451	1,04M	234,242	39,299

Table 1: Corpus sizes (techn.=technical)

3. Extracting paraphrases

In our previous study, we identified two types of potentially relevant paraphrases between technical and lay discourse types corresponding to two linguistic hypotheses. The first one, which is a common hypothesis (Fang, 2005), is that specialized language uses more nominal constructions where lay language uses more verbs instead. A second hypothesis is that medical language contains a fair proportion of words from Latin and Greek origins, which are referred to as neo-classical compounds. The meaning of these words may be quite obscure to non-experts readers. So one would expect to find less of these words in lay texts and instead some sort of paraphrases in common language. We therefore tried to detect paraphrases between nominal constructions in the technical side (such as *treatment of the disease*) and corresponding verb phrases in the lay side (such as *the disease is treated*), and paraphrases between neo-classical compounds (e.g. *gastritis*) and corresponding modern-language phrases (e.g. *stomach inflammation*). For French, the first kind of paraphrases proved interesting whereas the second type brought inconclusive results. We adapt here our extraction method for those two types so as to compare results and conclusions.

3.1. Pre-processing

As pre-processing steps, the two corpora were tokenized and segmented into sentences. Part-of-speech tagging was also performed using Treetagger. Our initial approach for French included an additional preliminary step consisting in segmenting the corpus into topic segments and in selecting only the most similar pairs of technical and lay segments according to a similarity measure (Cosinus). Here we took a slightly different stand and chose to look for paraphrases anywhere in the corpus, i.e. between any pair of technical and lay sentences. This decision was based on a comparative evaluation conducted for French which seemed to indicate that reducing the search to similar topic segment pairs was not conclusive.

3.2. Nominalisation paraphrases

This type of paraphrases involves nominalizations of verbal phrases and is built around the relation between a deverbal noun (e.g. *treatment*) and its base verb (e.g. *treat*). The general method was to look for corresponding content words (mainly nouns and adjectives) in the contexts of a pair of deverbal noun and corresponding verb. A set of extraction rules was thus defined, relying on a lexicon of deverbal nouns and verbs, a lexicon of synonyms, and a stemming phase. The English extraction rules were directly adapted from the original French rules. This mainly involved adjustments to take care of syntactic differences between French and English. Examples of adapted rules are given in Table 2. The left side of a rule represents a pattern to be found in the technical side and the right side a pattern to be found in the lay side. We also used WordNet[8] instead of lexicons to detect noun / verb pairs and synonyms. Approximately 20 rules were used to extract this type of paraphrases.

The patterns thus designed are close to the transformation rules of (Jacquemin, 1999) who detects morpho-syntactico-semantic variants of terms in plain monolingual corpora. One difference is that our patterns are built around one specific type of morphological variation (noun to verb variation) that seemed relevant in the context of the specialized/lay opposition, as opposed to any possible variation.

[2] http://www.ncbi.nlm.nih.gov/pubmed/
[3] http://www.merck.com/mmpe/index.html
[4] http://www.guideline.gov/
[5] http://medlineplus.gov/
[6] http://www.webmd.com/
[7] http://www.netdoctor.co.uk/
[8] http://wordnet.princeton.edu/

We also identify the paraphrases by comparing the two sides of a comparable corpus while (Jacquemin, 1999) starts from a given list of terms and searches for their variants in a plain monolingual corpus. Finally, we do not apply our method to terms specifically but to any expression corresponding to the patterns.

3.3. Paraphrases of neo-classical compounds

The second type of paraphrases we extracted were paraphrases between neo-classical compounds and their modern-language equivalents. In the French version of our method, we used a morphosemantic analyzer (DériF (Namer, 2009)) to detect neo-classical compounds in the technical side of the corpora and to decompose them into their constituent parts, each part being assigned its modern-language equivalent word (e.g. *gastrititis* = *gastr+itis* = *stomach+inflammation*). Simple search patterns to look for the modern-language components of the neo-classical compounds were then applied to the lay side. Here again, transposition of the French patterns to English was easily performed with a few syntactic modifications (see Table 3 for some examples). Detection and decomposition of neo-classical compounds was also performed through morphosemantic analysis, for which we used the English version of DériF (adapted from French in the context of another experiment (Deléger et al., 2009)). A total of 8 patterns were used.

3.4. Evaluation method

We examined the results of our experiment at two levels:

1. the precision of the extracted paraphrases;
2. the coherence of the initial linguistic hypotheses (hypothesis on nominalizations and hypothesis on neo-classical compounds).

3.4.1. Precision

We evaluated the quality of the extracted paraphrases by measuring their precision, that is, the percentage of correct results over the entire results. For the cancer corpus we evaluated precision on the whole results. In the case of the diabetes corpus, however, results being more numerous, we computed precision on a subset of 500 paraphrases (for each paraphrase type).

3.4.2. Coherence of the initial hypotheses

As explained above, this work relies on the following two linguistic hypotheses:

1. technical language tends to prefer nominalizations while lay language favours verbal constructions;
2. technical language tends to prefer neo-classical compounds while lay language favours modern-language equivalents;

To verify these two hypotheses, we relied on the paraphrases extracted with our method and computed frequency measures (in the same line as (Cartoni, 2008) for prefix alternation). More precisely, we looked at the proportion of a type of expression (nominalizations for the first hypothesis, neo-classical compounds for the second hypothesis) compared to possible cases (nominalizations+verbal constructions for the first hypothesis, neo-classical compounds+modern-language equivalents for the second hypothesis) in the technical side of the corpora and in the lay side. This proportion is referred to as the *preference index* of an expression. For each paraphrase identified through our method, we measured a preference index in each side (technical, lay) of the corpora. We then computed the mean of the indexes for each side of the corpora, so as to compare their values.

Formally, the preference index of a nominalization is computed as follows:

$$I = \frac{Nb_N}{Nb_N + Nb_V}$$

with Nb_N the number of occurrences of the nominalization and Nb_V the number of occurrences of the verbal construction.

Given our initial hypothesis that technical language uses more nominalizations whereas lay language uses verbal constructions, we would expect the preference index to be strong (close to 1) in the technical side of the corpora and low (close to 0) in the lay side of the corpora.

In the same way, the preference index of a neo-classical compound is written as:

$$I = \frac{Nb_C}{Nb_C + Nb_M}$$

with Nb_C the number of occurrences of a neo-classical compound and Nb_M the number of occurrences of its modern-language equivalent. As for nominalizations we expect the index to be high in the technical part of the corpora and low in the lay part of the corpora if our second hypothesis is confirmed.

4. Results

Table 4 shows the number of extracted paraphrases of each type, as well as their ratio per word of the corpora. We extracted many more occurrences of nominalization paraphrases than of paraphrases of neoclassical compounds.

	Diabetes		**Cancer**	
	Nb	*Ratio*	*Nb*	*Ratio*
Nominalization paraphrases	3,136	0.224‰	435	0.159‰
Paraphrases of neoclassical compounds	883	0.063‰	14	0.005‰

Table 4: Number of extracted paraphrases (types)

4.1. Precision

Table 5 shows that precision is rather good for nominalization paraphrases and that there is no major difference between the two corpora. As for neo-classical compounds, precision is average for the diabetes corpus and good for the cancer corpus, but this last figure is not significant given

	Paraphrases	Correct paraphrases	Precision
Diabetes	500 (sample)	370	74%
Cancer	435	337	77.4%

Table 5: Precision results for nominalizations

	Paraphrases	Correct paraphrases	Precision
Diabetes	500 (sample)	267	53.4%
Cancer	14	12	85.7%

Table 6: Precision results for neo-classical compounds

the limited number of paraphrases. These results are similar to those obtained for French: a precision ranging from 71.4% to 78.5% for nominalization paraphrases, and from 61.5% to 100% for neo-classical compounds. The number of paraphrases of neo-classical compounds is also very small for the smallest corpora. However the largest English corpus produced many more compound paraphrases than the French equivalent diabetes corpus. We attribute this difference to the restriction to similar text segments in our former French-language experiment.

Tables 7 and 8 give examples of extracted paraphrases for each type of paraphrase. The last lines of both tables show incorrect paraphrases, due to a synonymy link not valid in this particular context in the first case and to an only partial equivalence in the second case.

Technical	Lay
blood replacement	replaced blood
confirmation of diagnosis	confirm a diagnosis
absorption of insulin	insulin is absorbed
blood glucose fluctuations	blood glucose fluctuates
removal by surgery	removed in an operation
removal of the entire prostate	remove the whole of the prostate
gene carriers	carry a gene
*practice recommendations	exercise is recommended

Table 7: Example nominalization paraphrases (* = incorrect paraphrase)

4.2. Coherence of the initial hypotheses

Table 9 shows the mean preference index computed for each type of paraphrases in each side of the corpora. As expected, the index is high in the technical side and low in the lay side, for both paraphrase types. This seems to indicate that technical language has indeed a tendency to prefer nominalizations, while lay language favours verbal constructions, which was our first hypothesis. In the same way, technical language seems to use more neo-classical compounds than lay language which uses more modern-language equivalents, which corresponds to our second hypothesis. However, note that the smaller number of cases

Technical	Lay
pancreatitis	inflammation of the pancreas
haematuria	blood in the urine
erythrocyte	red blood cell
acidaemia	acid builds up in the blood
angiopathy	disease of the blood vessels
hyperglycaemic	blood has excess sugar
amylase	enzymes that digest starches
*normoglycemic	blood sugar levels rising above the normal

Table 8: Example paraphrases for neo-classical compounds (* = incorrect paraphrase)

for neo-classical compounds (only 12 in the cancer corpus) tempers this observation.

	Diabetes		Cancer	
	technical	*lay*	*technical*	*lay*
Nominalization paraphrases	0.84	0.23	0.77	0.07
Paraphrases of neoclassical compounds	0.98	0.28	0.84	0.14

Table 9: Mean preference index for each side of the corpora, and each type of paraphrase

5. Discussion

The results of this experiment are close to those obtained for French with our initial approach. This shows that a same simple rule-based technique can be applied to two languages such as French and English to extract paraphrases between two different discourse types. Furthermore, the adaptation of the method was performed with minimal effort and did not require a new corpus study, which definitely saved time.

This study also illustrates that relevant paraphrasing patterns in the context of the opposition between technical and lay discourse types are not necessarily language-specific. From the two paraphrasing patterns studied here, similar conclusions are drawn in English and French. That is, paraphrases between nominalization and verbal constructions seem to be an interesting direction as regards the technical/lay opposition. Paraphrases between neo-classical compounds and modern-language phrases are less relevant given their smaller number. However, results for these paraphrases are more numerous for English, due to the implementation of the approach. Indeed, while French paraphrases are obtained from pairs of similar text segments, English paraphrases are extracted from any sentence pair in the corpus, which consequently brought more paraphrases. Besides, there is no significant difference in precision, which advocates the use of an unrestricted search as we performed for English.

Finally, an originality of this work is that we use a comparable corpus of varied sources (the diabetes corpus), thus

containing rather dissimilar documents, as opposed to most existing approaches which often use closely related corpora (e.g. news corpora). We previously showed that this type of corpora could give good results for French. Here, we extend the conclusion to English, as we did not notice any significant difference in precision between the two corpora we used (the cancer corpus with closely related documents and the diabetes corpus with heterogeneous documents).

Future work includes applying the method again on French without restricting it to similar text segments, refining and increasing paraphrasing rules for nominalization paraphrases, identifying new patterns of relevant paraphrases between the two discourse types and testing less constrained paraphrase extraction methods (i.e. without necessarily predefining specific paraphrasing patterns to extract).

6. Conclusion

In this paper, we presented the adaptation of a method to extract paraphrases between technical and lay discourse types from French to English. Results show that paraphrases could be identified with a rather good precision, and that these types of paraphrase are relevant in the context of the opposition between technical and lay discourse types. These observations are consistent with the results obtained for French, which demonstrates the portability of the approach as well as the similarity of the two languages as regards the use of those kinds of expressions in technical and lay discourse types.

7. References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL*, pages 16–23, Edmonton, Canada.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 50–57, Toulouse, France.

Bruno Cartoni. 2008. Mesure de l'alternance entre préfixes pour la génération en traduction automatique. In *TALN*, Avignon, France.

Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-Parallel Corpora*, pages 2–10, Singapore. Association for Computational Linguistics.

Louise Deléger, Fiammetta Namer, and Pierre Zweigenbaum. 2009. Morphosemantic parsing of medical compound words: transferring a French analyzer to English. *Int. J. Med. Inform*, 78(Supplement 1):48–55.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *ACL BioNLP Workshop*, pages 49–56, Prague, Czech Republic.

Zhihui Fang. 2005. Scientific literacy: A systemic functional linguistics perspective. *Science Education*, 89(2):335–347.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing*, pages 57–64, Sapporo, Japan. Association for Computational Linguistics.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 341–348, College Park, Maryland.

Aurélien Max. 2008. Local rephrasing suggestions for supporting the work of writers. In *Proceedings of GoTAL*, pages 324–335, Gothenburg, Sweden.

Fiammetta Namer. 2009. *Morphologie, lexique et traitement automatique des langues : l'analyseur DériF*. Lavoisier, Paris.

Marius Pasca and Peter Dienes. 2005. Aligning needles in a haystack: Paraphrase acquisition across the web. In *Proceedings of IJCNLP*, pages 119–130.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing (IWP)*, pages 65–71, Sapporo, Japan. Association for Computational Linguistics.