

ONTOLOGY & LEXICON

Final report

Ingrid Falk, Charif Haydar, Nadiya Yampolska
Supervisors: Paul Buitelaar, Claire Gardent

as part of Natural Language Applications course, DFKI / LORIA

January 26, 2008

Contents

1	Project Description	2
1.1	Motivation	2
1.2	Goals and objectives	2
1.3	Related work conducted at DFKI	2
1.4	Project data	3
2	Methodology	5
2.1	Original WSD synset-based algorithm	5
2.2	Mann-Whitney test	7
2.3	Single synonym based χ^2 test	8
2.4	Implementation	9
3	Results and evaluation	9
3.1	Evaluation benchmark: human judgments	9
3.2	Synset-based WSD algorithm	10
3.3	Mann-Whitney test	11
3.4	Single term based χ^2 test	13
4	Discussion	14
4.1	Challenges	14
4.1.1	Resources for languages other than English	14
4.1.2	Evaluation benchmark	15
4.1.3	Using synonym dictionary	15
4.2	Perspectives	15
A	Overview of the scripts	16
B	Result Summary	17

1 Project Description

1.1 Motivation

The project was inspired by the research done in German Research Center for Artificial Intelligence by Nils Reiter and Paul Buitelaar on enrichment of human anatomy ontology by lexical information. Domain specific ontologies are more and more important for effective natural language processing and applications. Concepts in a domain specific ontology are organised in hierarchical structures describing very rigorously a particular branch of knowledge. Ontologies aim at providing very specific information, with a unique term corresponding to one concept linked to other concepts by different relations. Therefore ontologies, providing a great amount of detail about concepts and relations, become a powerful resource for applications, e.g. semantic search, text understanding, etc. The obstacle for using the full potential of ontological knowledge is in fact the nature of their construction, namely, providing a very specific unique term per concept. Natural language, in contrast, also by its nature is a very flexible and varied tool to point to the same concept in a number of ways. If ontologies could be enriched by the *lexical information*, information about how ontological concepts are referred to and used in language, it could open opportunities to exploit this knowledge in a much broader sense.

1.2 Goals and objectives

The goal of ONTOLOGY & LEXICON (ONTOLEX) is lexical enrichment of a domain specific ontology in human anatomy through automatic word sense disambiguation. Since ONTOLEX is a project closely related to research done for the English language working with the same ontology, the parallel goal of the project is providing multilingual lexical information for ontological concepts, namely, for the French language.

Our objectives are:

- test and evaluate the lexical enrichment algorithm through word sense disambiguation on another language rather than English;
- explore the opportunities for exploitation of available data for the French language;
- outline the challenges of the task for French and suggest other ways to improve and build on the original algorithm.

1.3 Related work conducted at DFKI

The algorithm for lexical enrichment through word sense disambiguation was suggested by [5]. In a nutshell, the algorithm uses the WordNet synonym set information to provide synonyms for an ontological term, and synonym frequency

information in a domain specific versus general corpus to identify the correct synonym set. The idea behind the approach is that manually constructed synonym sets in WordNet are strongly related by meaning and selecting a domain specific sense of a term would allow to automatically enrich a concept with all its lexical synonyms in this set. The task of finding a domain specific sense of a term is handled by a χ^2 test, a widely used test for evaluation of statistically significant differences between proportions for two or more groups in a data set.

The data which [5] used for their experiment is summarised below:

Base ontology	Foundational Model of Anatomy Ontology ¹ ([6])
Ontology size	75,000 classes and over 120,000 terms
Linguistic resource	WordNet 2.0[3]
Domain corpus	Wikipedia category articles on Human Anatomy
Size of domain corpus	4,4 million words
Reference corpus	Frequency lists of British National Corpus

The intersection of WordNet synset data and ontological concepts gave 1,382 terms for English. For these terms, the word sense disambiguation (WSD) algorithm was applied, selecting the synset with the highest accumulative χ^2 score. Other WordNet relations like meronymy, hypernymy and hyponymy were also incorporated in the calculations, which positively affected the results from 59,70% to 73,63% f-score.

1.4 Project data

As mentioned previously, we have chosen to work with the same ontology FOUNDATIONAL MODEL OF ANATOMY (FMA). The main reason for this choice is that FMA comes with French terms for many of its concepts and in the case of success we would be able to provide multilingual lexical support for the ontology covering two languages. The total number of FMA French terms is **4,5 thousand words**.

As our LEXICAL RESOURCE we have chosen to work with the synonym dictionary *Le Petit Robert* whose intersection with the FMA terms is **75 distinct terms** and whose synonym structure resembles synonym sets in English WordNet. We found it most complete and suitable for our tasks. Other synonym dictionaries considered were Bailly, Benac, Duchaz and Larousse. In addition, it is a rather recent dictionary, so we may expect it to feature more domain specific terms. The idea of using *French WordNet* was put aside, since the number of ambiguous terms in French WordNet corresponding to FMA terms gave only 34 terms, and many of them not providing any synonyms. The ambiguity of these terms was indicated by enumeration of various senses of a term, but disambiguation of the senses would be impossible without accompanying synonyms.

The DOMAIN CORPUS was automatically crawled from various sources. One of the obvious sources was Wikipedia articles exported from the category Cat-

egory::medecine and Category::anatomie_humaine. The resulting number of words extracted from Wikipedia was **9,969 tokens**. As this was not enough for reasonable calculations, we used the **CISMef** index to extract web sites related to human anatomy.

CISMef is a French online catalog and index of Health Internet resources ([1]). The site provides a hierarchically organised thesaurus of medical terms. One of the branches of the tree is **Anatomie**, which has the following subterms: *régions du corps, appareil locomoteur, système digestif, appareil respiratoire, appareil urogénital, système endocrine, système cardiovasculaire, système nerveux, organes des sens, tissus, cellules, liquides et sécrétions biologiques, anatomie animale, système stomatognathique, systèmes sanguin et immunitaire, structures embryonnaires, système tégumentaire*. Apart from *anatomie animale* we found all these categories relevant to our domain collection. We have crawled the sites indexed by these terms for .html, .pdf or text documents and converted them to plain text. The majority of documents were university courses in physiology (CHU d'Angers), histology (CHU Pitié Salpêtrière), pathological anatomy (Université Paris 7), embryology (Université Strasbourg), and others (for full list, please address GForge - Domain corpus - Links). The resulting domain corpus made up **1,6 million words**.

We have tagged the corpus with the **TreeTagger** and produced frequency count list for the ontological concepts which had an intersection with Le Petit Robert synonym dictionary.

We took **Frantext** general corpus of novels dating 1950-2000 and their frequency lists as our REFERENCE CORPUS which was available from the LORIA Talaris group. The total size of the reference corpus is **29 million words**. The choice of Frantext was very appropriate for our purposes since we do not expect fictional novels to contain medical or anatomical terms, unless of course it is 'La Peste' by Albert Camus.

The corpus had been previously parsed and annotated with **Syntex** and is available in the .xml format. We used the corpus in the given format to produce frequency count list for the ontological concepts.

The summary of the project data is provided below:

Base ontology	Foundational Model of Anatomy: 4,575 terms
Linguistic resource	Le Petit Robert synonym dictionary
Domain corpus	A manually collected corpus using the CISMef : Index of French language Health Internet resources [1].
Reference corpora	Frantext 3.01: 15 mill. words, 1950-2000, a collection of literary texts

2 Methodology

During the project, we have tested the original word sense disambiguation algorithm used by [5], adapted it to better match our data introducing normalisation and threshold, and suggested other ideas about how to improve the results².

2.1 Original WSD synset-based algorithm

implemented by Nadiya Yampolska

The method applied was described in detail in [5]. Each ontological concept is looked up in a lexical resource like WordNet, or Le Petit Robert in our case, providing information about one or more senses of the term used in different contexts. The χ^2 score is calculated from the contingency matrix of each term and each synonym in the domain corpus and the reference corpus. If the term is ambiguous, i.e. it has more than one synonym set, a cumulative χ^2 score for the synset is calculated and the synonym set with the highest ranking is selected as a synset corresponding to the domain sense of the term. We have calculated the χ^2 score for each synonym with each possible tag for the term from the TreeTagger French tagset. We have restricted the possible tags to the ‘content’s tags, i.e. Adjective, Noun, Adverb and Verb of different types, avoiding counts of prepositions, determiners, abbreviations, etc.

Provided the ontological term *estomac* and the following synonym sets from Le Petit Robert dictionary:

[1] *estomac*: *aplomb*, *cran*, *culot*

[2] *estomac*: *diaphragme*

[3] *estomac*: *tripes*, *ventre*

is expected to select sense [2]: *estomac*: *diaphragme*.

The χ^2 score for sense [1] is the sum of χ^2 scores for *aplomp*, *cran* and *culot*, which returns approximately 55.75. The sense [2] itself scores 758.18, showing a big difference with respect to domain specific presence compared to sense [1]. In the synset [3], the word *tripes* does not occur in either domain or reference corpus, therefore its presence in the synset does not contribute to the system decision. The χ^2 score for *ventre* is 332.84.

The resulting ‘best scoring’ synset is sense [2], as expected.

After the selection of the synset, all the synonyms of the synset are added to the ontological concept under the assumption that the synonym sets are grouped by closely related meaning and are used in similar contexts. This assumption

²The scripts performing each method described in this section can be accessed at GForge

heavily depends on the data, and we discuss the related challenges in section 4.

The method described above is an exact copy of the method used by [5] tested on new data. We have proposed several amendments to the approach after analysing the results of the original method (all results are accessible at GForge and discussed in section 3). The suggestion was to normalise the χ^2 cumulative result based on how many terms have contributed to the total score. Under the assumption that if a synset containing a single term scores high, we should prefer it to the synset containing five or six different terms with the same score, we do so by dividing the synset score by the number of synonyms which contributed the non-null score. The motivation for taking such average score is intuitive. Let us assume that synset A containing 5 synonyms has scored 15 total, whereas synset B containing 1 synonym scored 13. What synonym set is more likely to represent domain specific data: A or B? Since we take the sum of all synonyms, we can see that even though synset A obtained the highest rank, it consists of 5 terms each having a relatively low rank. On the other hand, synset B has one synonym which on its own has a considerable weight. In this case, we prefer to select synset B as the best domain specific sense of a term.

Not all terms contribute to the total weight of the synset. As we have seen, the term *tripes* did not occur in either domain or reference corpus and therefore was discharged from synset χ^2 calculation. We keep track of such instances and only divide the total synset *chi*² score by the number of synonyms which had a non-zero weight.

Using such normalisation has resulted in slight improvement in the selection.

Another novelty in our approach is the introduction of a threshold for making a selection decision. The original approach *always* selects a synset. This of course is effective if we assume that there is always a sense of the ontological term which points to domain specific usage in our lexical resource. Since English WordNet is a very rich resource, we can expect this to be true. However, working with a smaller range synonym dictionary or other less detailed resource, we come across the problem that there may be no appropriate synonym set for the term. In this case, selecting a best synset may evoke false results. Indeed, choosing between the following synsets for the term *langue*

[1] langue: bavarde, dard, lavette, menteuse

[2] langue: idiome

[3] langue: langage

[4] langue: languette

[5] langue: bavarde, bec, bouche, discours, langage, menteuse, parole

we expect the system to return the negative result, i.e. **no synset belongs to domain related sense of the term *langue***. We do so by introducing a χ^2 threshold, defined experimentally, below which our system recommends to not add any synonyms to the ontological concept due to low scores.

The results of all three variations of the original method and their comparison table can be accessed at GForge and discussed in section 3.

2.2 Mann-Whitney test

implemented by Ingrid Falk

After analysing synonym dictionary data more closely we came to the conclusion that the synonyms are not always well grouped for our purposes. One of the indications for this was the difficulty we had in establishing our own judgments about what synonym sets belong to human anatomy domain. For more details, see section 3.1. We have decided to get a more precise information on how the distributions of the ontological terms and their synonyms differ in the domain specific corpus as compared to the reference corpus. This approach is discussed in [4], where Adam Kilgariff recommends to use the Mann Whitney test for such task.

The Mann-Whitney rank sum test (also known as the Wilcoxon rank sum test) is a non-parametric test for assessing whether two samples of observations come from the same distribution. The test can be applied to corpus data, if two corpora are first divided into same-sized samples. The null hypothesis is that all the samples are from the same population. We count the occurrences of a given word in every sample and then rank the samples from both corpora using this frequency information. Thus, the problem is reduced to the distribution of rank sums of data sets of different size. This, in its turn, has been previously worked out, and the distribution tables are freely available.

The Mann-Whitney test gives an indication of:

- whether the difference of the observed and expected rank sums may be due to chance (with confidence 99.5)
- whether the counts from the domain specific corpus are
 - *bigger* than expected (with confidence 99.5)
 - *smaller* than expected

We have applied the test on every term and each of its synonyms. To do so, we divided each corpus into chunks of 100,000 words. For each ontological term or its synonym, we computed the frequencies in each of the chunks. As a result, we get two data sets — one for each corpus — with frequency counts. The counts are ranked and the sum of the ranks for each of the data sets is computed. The Mann-Whitney test returns the probability of these rank sums given the size of our data sets: if the size of the data sets is n_1 and n_2 resp. and $N = n_1 + n_2$, the expected rank sum W_1 of the first data set is $W_1 = n_1 * (N + 1)/N$ under the assumption that the data is equally spread over the 2 data sets. The probability can be computed directly for small data sets ($N = 20/30$), and by normal approximation for larger N . The test works for $N \geq 5$.

The results of Mann-Whitney test are summarised and discussed in the section 3.

2.3 Single synonym based χ^2 test

implemented by Charif Haydar

This section discusses two experiments based on single synonym scores. We discuss why these scores can be important given our data and how they can be used. The first experiment aims at defining new synonym sets based on human anatomy usage via χ^2 scores, whereas the second introduces a different measure for indication of significance of a term in a specific domain.

Having noted that the synonyms in Le Petit Robert are not always well grouped for our purposes, we have decided to also perform the χ^2 test on single synonyms as opposed to the whole synset. In other words, since the synonyms are grouped in a synset by criteria other than domain of usage, by making our judgment on the whole synset we a) risk to eliminate a synonym which does in fact belong to the domain of human anatomy, and b) incorrectly add a synonym with a very low χ^2 score to the ontological concept because of the group it belongs to. **Note, that our assumption here is that the synonym sets in Le Petit Robert are grouped based on some different criteria and do not reflect the domain of usage.**

This third module of our system is based on χ^2 scores of single synonyms. In our setting, when using the χ^2 test the line of reasoning put forward in statistic textbooks would be the following: the null hypothesis is that both corpora comprise words drawn randomly from some larger population. Then, whenever the χ^2 statistic is greater than the critical value of 7.88, we conclude with 99.5% confidence that, in terms of the word we are looking at, the two corpora are not random samples of the same larger population.

We computed the χ^2 scores for each term and its synonym. For a great majority of terms, the score is much bigger than the critical value, so in this form the test was of little use. However, this confirms the results of A. Kilgariff in [4].

We tried to adapt the method by setting an experimentally defined threshold. Each synonym surpassing this threshold is added to a new “usage based” synset, independently from which original synset the synonym was taken. This “usage based” synset represents the synonyms which finally should be added to the concept.

The threshold and obtained synonym sets are discussed in section 3.

For the next experiment we have used a less complicated formula than χ^2 , but from the same family and with the same underlying idea. We consider a

certain synonym to belong to the anatomy domain when its frequency in this domain is reasonably large, at the same time when it is not a ‘general purpose term’. In our new measurement, we subtract the frequency of a term in the reference corpus from its frequency in the domain, which indicates a very simple difference measure of its usage in these corpora. We also take into consideration the size of each corpus, so each term frequency is divided by the corpus size from which it was calculated. In addition, to encourage large frequencies to have more effect than small frequencies, we square the frequency values in both corpora as well as eliminating synonyms which did not occur in either corpora. The resulting formula which we used for this experiment is the following:

$$\frac{\text{DomainFrequency}^2}{\text{DomainSize}} - \frac{\text{ReferenceFrequency}^2}{\text{ReferenceSize}}$$

The results again are discussed in section 3.

2.4 Implementation

All methods and all pre-processing steps (web crawling, frequency counts, etc.) are implemented in Perl. The system receives text files as input, and the main output files are .html files containing result tables for presentation purposes.

We have kept track of all our discussions, options, data and experiments using the Inria Gforge development service ([2]). All implementation scripts, intermediate and final result files, documentation, reports, evaluation benchmark, multimedia presentation (presented in November 2007 at Université Nancy 2) are available at ONTOLEX URL (<https://gforge.inria.fr/projects/ontolex/>)³.

Also see appendix A for an overview of the scripts.

3 Results and evaluation

3.1 Evaluation benchmark: human judgments

In order to evaluate the results of our system, we have assessed each synonym set of each ontological term as for its relevance to the domain of human anatomy. Three of us have done so independently from each other, and then merged our judgments. As a result, out of 225 unique synsets, 73 synsets were labeled with at least one positive judgment. The three possible judgments were ‘yes’, ‘no’ or ‘I don’t know’ wrt. can the synonyms of this set be considered as belonging to the domain of human anatomy.

The task was very difficult for a number of reasons. First of all, none of us is a native French speaker. We had to translate many of the terms, which in itself results in a lot of ambiguity and depends heavily on what source each of us

³for project members only

used to do the translation. One of us has worked with a French-fluent medical student to make the judgments, however, it was still difficult to make a final decision.

On top of the language barrier, the synonym groups did not seem to represent a single well determined sense, sometimes grouping together the clearly medical or anatomical terms with very general ones. For example, the synset *compartiment*: **alvéole**, *box*, *case*, *casier*, *cellule*, *chambre de chauffe*, *loge*, *stalle* where the term **alvéole** is clearly an anatomical term, but the rest is hardly related to the same domain.

The bottom line is that making the judgments was very hard, and unfortunately we cannot be sure about their accuracy. However, the inter-judge agreement was still very high. The summary of the agreement is provided below. ‘Unknown judgments’ refer to those where the synset was labeled by ‘yes’, ‘no’ and ‘I don’t know’ at the same time. Agreed judgments refer to both positive and negative labels.

Total judgments	226
Agreed by 3 judges	165
Agreed by 2 judges	54
Agreed <i>at least</i> by 2 judges	219
Unknown judgments	7
Judgment agreement in percentage	96.90%

3.2 Synset-based WSD algorithm

We have run the WSD original algorithm on every synset that a) had at least one positive human judgement, and b) was not the only synset for its term. We have calculated the cumulative χ^2 score for each of them and selected the best scoring one. We have also amended the algorithm by normalising the total χ^2 score by the number of synonyms which made the χ^2 contribution to the sum. After doing so, we have also introduced a threshold which separates the recommendation of our system to include the synset to the ontology or not. More detailed description is provided in section 2.

The complete results can be accessed at GForge. As a summary of the obtained results, we can conclude that the χ^2 best score WSD algorithm did not work very well on our data. This can be for several reasons. One is that the synonym sets were not very clear cut. The second perhaps is a reflection of our corpora size: 1,6 million for domain corpus and 29 million for reference corpus. Perhaps, having more domain specific data would allow the system to make better predictions. The domain specific material can also be an issue and could seriously influence our results. There was no perfect distinction made between ‘medical’ domain and ‘human anatomy’ domain, which could again have its consequences.

The summary of the results for the synset-based WSD algorithm is given below. Note, that we consider our recall of 100% since our system **always** returns a result, at the expense of precision.

	Best-score	Norm. best-score	Threshold ≥ 500
How many correct	18	20	19
How many total	41	41	41
Precision	43.90%	48.78%	46.34%
Recall	100%	100%	100%
F-score	32%	36%	34%

As is clear from the table above, the results of all three modifications of the original algorithm are close to each other. However, we do have a 5% increase in precision when normalising the χ^2 score. A vivid example of such improvement would be the term *organe* which out of its six synsets has two which were marked by positive labels: [1] ‘corps’, and [2] ‘membre, penis’. The best score algorithm selects the synset ‘organe: agent, centre, instrument, moteur’ which is in fact the synset with the largest number of synonyms. The normalised score algorithm selects the synset [2] instead, which is one of the expected synsets. It must be mentioned though that in the majority of cases the system decision does **not** change after normalisation, and if changing, does not always return the better solution. It is however an issue to consider when dealing with a cumulative result with a different number of ‘participants’.

An example where threshold plays a positive role would be, among others, the following. Let us again consider the term ‘langue’ and its synonym sets, which has the total of 5 of them. Both best and normalised χ^2 score algorithm returns the same synset, noticeably with the largest number of synonyms: ‘langue: bavarde, bec, bouche, discours, langage, menteuse, parole, platine, tapette’. It is at the same time competing with single synonym synsets, which could be to their advantage if they were not so hopeless: ‘idiom’, ‘languette’, ‘language’. However, the cumulative score of these 9 synonyms only return the χ^2 score of 224.576661917642. This is below our threshold and the system recommends not to add any of these synonyms to the ontological concept.

Although the results are only slightly different, we consider it necessary to have some sort of normalisation related to the size of a synset as well as provide the system with the opportunity to reject all synsets of a term as opposed to **always** suggesting the best candidate for lexical enrichment, as discussed in [5].

3.3 Mann-Whitney test

We applied the Mann Whitney test on every term and each of its synonyms. The test shows not only whether the difference in the distribution of the counts is significant, but also if the counts in one sample are usually bigger than in the other. Table 1 shows sample output for the terms in a synonym group of *bouche*.

Term	Domain specific corpus	Reference Corpus	Test outcome
entrée	–	+	significant
gueule	0	2057	
orifice	+	–	significant
ouverture	+	–	significant
embouchure	1	72	not significant

Table 1: Mann Whitney test results on the terms in a synonym group of **bouche**

The test suggests that the distribution of *entrée* is significantly different in the domain specific corpus from the one in the reference corpus, it tends to appear more often in the reference corpus chunks. *orifice* and *ouverture* also have significantly different distributions in the two corpora, but their frequency counts are usually bigger in the domain specific corpus. *gueule* does not occur in the domain specific corpus, thus the test is not applicable. The distribution of *embouchure* is not significantly different, it occurs once in the domain specific corpus and 72 times in the reference corpus.

This synonym group also was selected by the synset based χ^2 disambiguation method. Using the Mann Whitney results we could narrow the set of synonym candidates to *orifice* and *ouverture*. In this case the outcome looks plausible: *entrée* and *embouchure* (and *gueule*) probably are no “human anatomy” terms, whereas *orifice* presumably is one. As for *ouverture*, its distribution penchant to the “human anatomy” domain is plausible, but it does not seem right to consider it a synonym of *bouche* in this context.

In table 2 we show the test results for the best scoring (in terms of the synset based χ^2) synonym group of *compartment*, another instance where the method would perform acceptably:

Term	Domain specific corpus	Reference Corpus	Test outcome
alvéole	+	–	significant
box	0	81	
case	51	363	not significant
casier	0	195	
cellule	+	–	significant
chambre de chauffe	Neither in domain, nor in general corpus		
loge	56	655	not significant
stalle	0	65	

Table 2: Mann Whitney test for the best scoring synonym group of **compartment**

The following example illustrates another possible use case of the test. Table 3 shows the statistics for the best scoring synonym group of *langue*: *bavarde*, *bec*, *bouche*, *discours*, *langage*, *menteuse*, *parole*, *platine*, *tapette*. Our judgements disagree wrt. to this synonym group: one asserted this synonym group would

rather represent a “human anatomy” meaning, two not. As the results in table 3 show, the test would rather confirm the **no** judgement.

Term	Domain specific corpus	Reference Corpus	Test outcome
bavarde	Neither in domain, nor in general corpus		
bec	11	789	not significant
bouche	–	+	significant
discours	1	1341	not significant
langage	14	1247	not significant
menteuse	Neither in domain, nor in general corpus		
parole	–	+	significant
platine	1	12	not significant
platine	8	52	not significant
tapette	0	23	

Table 3: Mann Whitney test for the best scoring synonym group of **langue**

However, for the synonym group *myocarde: endocarde, péricarde* where we all agreed that it should represent a “human anatomy” meaning, the method would return the results shown in table 4.

Term	Domain specific corpus	Reference Corpus	Test outcome
endocarde	16	1	not significant
péricarde	26	0	

Table 4: Mann Whitney test for the best scoring synonym group of **myocarde**

Therefore, although the results look interesting, further investigation and reliable benchmarks are needed for more sound standing assertions.

3.4 Single term based χ^2 test

Let us consider the results obtained by the two experiments with the single term χ^2 scores. One of the main conclusions from these tests is that synonym sets in Le Petit Robert were indeed not well grouped for our purposes. In other words, if we assess the χ^2 values of individual synonyms in the same synset, we can see a great variation of these values.

To evaluate these results we consider the human positive judgement about a synset as a ‘go’ to **all synonyms in that synset** to be added to the ontology. Therefore, we label each synonym in the synset with a positive ‘yes’ judgement. The table below shows the χ^2 based decision about individual synonyms with different thresholds, and the level of disagreement with the human judgement is apparently very high. In fact, it reaches 50%. We think that this shows that the experiment deserved to be done for such illustration.

Threshold	Overcoming threshold	Disagreement with human judgement
50	230	300
100	173	283
150	143	281
200	126	272
300	90	273
500	64	271

The most balanced threshold in terms of how many synonyms it selects as anatomy-related synonym and how low the disagreement with the human judgement, is 200. Under such threshold, 74 out of 136 total synsets containing more than one synonym (54.4%) were ‘split’: the synonyms in the same synset were both below and above the threshold.

The following table shows the results using the ‘simplified’ formula with the same underlying idea as χ^2 test built on various values of a threshold, defined experimentally. The large number of disagreement with the human judgements is the indication, among possible others, that it is indeed true that synsets were not well grouped in respect to the domain usage. Individual χ^2 scores vary greatly among the synonyms of the same synset on the same domain and reference data, and 29.4% of synsets with more than one synonym was ‘split’ by this test.

Threshold	Overcoming threshold	Disagreement with human judgement
0	62	254
-0.001	222	300
0.001	33	273

If we compare the results of this simplified formula with the threshold of zero the results of χ^2 test, we can see that we managed to reduce the number of disagreements between human and system’s judgment, at the same time leaving a reasonable number of accepted synonyms. Needless to say, this method would need to be tested on other data.

4 Discussion

4.1 Challenges

4.1.1 Resources for languages other than English

One of the main challenges in all NLP applications for languages other than English is the incomplete resources. In our case, it was the French WordNet which could not serve us the same way as English WordNet could serve [5]. The previous study showed that using the hypernymy, meronymy and other WordNet relations was very useful for identifying the best synset. Since we were short of such resources, we had to resort to only using synonym information. Moreover, the synonym information which was available to us was not as fine grained as WordNet data.

4.1.2 Evaluation benchmark

To make a judgment about the whole synonym set whether or not it belongs to the domain of human anatomy is a difficult task. It is difficult for non-native speakers of French, it is difficult for non experts in human anatomy, and it is just genuinely difficult because knowing a word is different from knowing how it is used in a domain specific context. In addition, terms which are ‘medical’ are not necessarily the ones considered in the domain of anatomy in general, and human anatomy in particular. This was a big challenge for us, and we believe for anyone who is setting out for the similar task.

4.1.3 Using synonym dictionary

Unless one is sure about the criteria under which synonyms were originally grouped in one sense in a dictionary, it is difficult to say how such a synonym set is useful for one’s purposes. In our case, it was obvious that synonyms were not linked through domain related principles. The results discussed in Single Term χ^2 experiments point out that the domain specific usage of synonyms in the same synset were on the two sides of the threshold in more than 50% of cases. Therefore, identifying a ‘better’ synset would be possible using a χ^2 score test, but within that synset a different method (for ex., Mann-Whitney test) would be necessary to select the best synonym(s) to be added to the ontology.

4.2 Perspectives

We see the perspectives of combination of several methods for a) selection of a synset; b) selection of the best synonym(s) within the synset (it could be as well all of them); and finally c) defining some educated threshold under which the system would reject to add synonyms to the ontological concepts.

Another improvement of the results is possible with the increase of corpora sizes, and further exploration of lexical resources available for the French language.

Acknowledgments

We would like to thank Pierre Zweigenbaum for his very helpfull advice regarding the domain corpus in Human Anatomy. We are also grateful to our supervisors Paul Buitelaar and Claire Gardent for their support, and Nils Reiter for his collaborative remarks.

References

- [1] CISMeF: Catalog and Index of French-language Health Internet resources. A quality-controlled subject gateway.
- [2] The Inria Gforge, a service to assist scientific collaboration and development at Inria.
- [3] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1999.
- [4] Adam Kilgarriff. Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133, 2001.
- [5] Nils Reiter and Paul Buitelaar. Lexical enrichment of a human anatomy ontology using wordnet. In *WordNet08*, 2008.
- [6] C. Rosse and J.L.V. Mejino Jr. A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, 2003.

Appendices

A Overview of the scripts

FMA and “Le Petit Robert” terms

`fma_fwn_lookup.pl` Lookup words and counts in the French WordNet.

`fma_syno_base_lookup.pl` Lookup words and counts in the synonym bases.

Domain specific corpus

`wget.pl` Crawling the sites and collecting html, pdf and text files.

`wget2text.pl` Converting html and pdf to plain text.

`collect_corpus.pl` Collects the text files in one directory, counts, produces a content file.

`make_domain_corpus_chunks.pl` Makes the 100.000 word chunks needed for the Mann Whitney test.

`domain_corpus_frequency.pl` Counts word frequencies (of the FMA terms and synonyms).

Reference corpus

`extract_frantext_sax.pl` Extracts relevant data from the corpus

`make_frantext_chunks.pl` Makes the 100.000 word chunks needed for the Mann Whitney test.

`reference_corpus_frequency.pl` Counts word frequencies (of the FMA terms and synonyms).

The Tests

`xci-score_reportscript.pl` Computes χ^2 -scores for terms and synonym groups.

`single_term_chi_test.pl` Computes χ^2 -scores for terms.

`single_term_mod_chi_test.pl` Computes scores for terms, using a χ^2 variant.

`Statistics::Test::WilcoxonRankSum` Perl module to compute the Mann Whitney Rank Sum test on two sets of numeric data. This module is uploaded on CPAN (Comprehensive Perl Archive Network⁴).

`mann.whitney.pl` Does the Mann Whitney test on the terms from the FMA and *Le Petit Robert*.

`synsets_summary.pl` Computes the final results html file.

B Result Summary

χ^2 scores for the synonym groups http://ontolex.gforge.inria.fr/syngroup_chi_scores.html

χ^2 scores for the synonym groups, compared to benchmarks http://ontolex.gforge.inria.fr/results_evaluation.xls

χ^2 scores for single terms http://ontolex.gforge.inria.fr/single_term_chi_test.html

scores for single terms, computed using a χ^2 variant http://ontolex.gforge.inria.fr/single_term_mod_chi_test.html

Summary of the Mann-Whitney and χ^2 statistics for synonym groups and single terms http://ontolex.gforge.inria.fr/fma_robert_synsets_summary.html

⁴<http://search.cpan.org/~ingrif/Statistics-Test-WilcoxonRankSum-0.0.3/>