Ambiguity Diagnosis for Terms in Digital Humanities

Béatrice Daille[1], Evelyne Jacquey[2],
Gaël Lejeune[3], Luis Felipe Melo[4], Yannick Toussaint[4]
[1]LINA UMR 6241, University of Nantes; [2]UMR 7118 ATILF CNRS Université de Lorraine;
[3]GREYC UMR 6072, Normandy University; [4]INRIA Nancy Grand-Est & LORIA CNRS Université de Lorraine.
beatrice.daille@univ-nantes.fr, evelyne.jacquey@atilf.fr,
gael.lejeune@univ-nantes.fr, luisfe.melo@gmail.com, yannick.toussaint@loria.fr

Abstract

Among all researches dedicating to terminology and word sense disambiguation, little attention has been devoted to the ambiguity of term occurrences. If a lexical unit is indeed a term of the domain, it is not true, even in a specialised corpus, that all its occurrences are terminological. Some occurrences are terminological and other are not. Thus, a global decision at the corpus level about the terminological status of all occurrences of a lexical unit would then be erroneous. In this paper, we propose three original methods to characterise the ambiguity of term occurrences in the domain of social sciences for French. These methods differently model the context of the term occurrences: one is relying on text mining, the second is based on textometry, and the last one focuses on text genre properties. The experimental results show the potential of the proposed approaches and give an opportunity to discuss about their hybridisation.

Keywords: Terminology, Disambiguation, Ambiguity, Polysemy, Text Mining, Textometry, Salience

1. Introduction

This paper introduces and compares three original methods for ambiguity diagnosis. The objective is to decide for any occurrence of a term candidate (TC) if it is terminological (TO) or not (NTO). This ambiguity diagnosis is useful for information retrieval, keyphrase extraction, and also for text summarization. While lexical disambiguation is a very productive issue (Navigli, 2009), research on term disambiguation remains surprisingly unexplored. Nevertheless, in any domain, TCs may be ambiguous (L'Homme, 2004), having TOs and $NTOs$ as well.

As an illustration, let us consider two occurrences of *aspect* in a research paper belonging to the linguistic domain in French :

> (I) *L' `aspect` est une catégorie qui reflète le déroulement interne d'un procès* 'Aspect is a category which expresses the internal sequence of a process' (Cothire-Robert, 2007)
>
> *(II) Ce dernier `aspect` est primordial* 'This last aspect is primordial' (El-Khoury, 2007)

In the first example, `aspect` is a term, while it is not in the second example. Ambiguity occurs also for multiword terms when they are submitted to lexical reduction in discourse. The reduced form is often ambiguous: the occurrence of the TC `analysis` could refer to `syntactic analysis`, `semantic analysis` or to a non-terminological sens of `analysis`.

Term disambiguation, as in most works on word sense disambiguation, is considered here as a classification problem. We propose three supervised learning methods which explore different modelizations of TCs context. They are tested on a manually annotated corpus in the broad domain of social sciences in French.

This paper is organised as follows: first we give some insights about related work (section 2.), then we present the dataset (section 3.) and the proposed methods (section 4.). Finally, we evaluate the methods (section 5.) and discuss their results (section 6.).

2. Term Desambiguation versus Term Acquisition

Our term disambiguation process comes after the automatic term acquisition (ATA) task. Indeed, ATA tools extract TC considering criteria and indices computed over a whole corpus. Thus, they take a global decision for the TC. If a string is identified as a TC, all its occurrences are considered as TOs. The diagnosis task is slightly different since a decision is taken for each TC occurrence.

Several machine learning methods have been used for ATA. Foo and Merkel (2010) propose RIPPER, a rule induction learning system that produces human readable rules. Potential terms are n-grams, mainly unigrams, occurring in Swedish patent texts. The features that have been used are linguistic features such as POS tags, lemmas and several statistical features. The best configuration gave 58.86% precision and 100% recall for unigrams. Judea et al. (2014) used CRF on TCs occurring in English patent texts. TCs that are submitted to the classifier satisfy syntactic patterns. They developed a set of 74 features that include the POS tags of the TCs, their contexts (adjacent bigrams), corpus and documents statistics and patents properties. The best configuration gave 83.3% precision and 74.3% recall.

3. Dataset

The dataset contains 55 documents (13 journal papers, 144,000 tokens), and 42 conference papers, 197,000 tokens) from the SCIENTEXT corpus (Tutin and Grossmann, 2015). The TERMSUITE tool (Daille et al., 2011) has been used to lemmatize the corpus, POS tag and extract TCs. Any other term extractor could have been used and the comparison of the performance of term extractors is out of the scope of this paper. The resulting data is the benchmark for a cross-validation approach. Each TC occurrence has been manually annotated as TO or NTO by an expert following a four step annotation process (Gaiffe et al., 2015). 4,204

TCs have been extracted corresponding to 52,168 occurrences. Only 33.10% of these occurrences and 35.10% of TCs have been annotated as TOs by the experts. To facilitate the annotation process and to make it as expert independent as possible, the task has been divided into four manual disambiguation (MD) steps. Each step corresponds to a MD_i label : experts are asked for MD_{i+1} annotation only if MD_i is positive.

For each occurrence of a TC, the expert should answer if: (1) it is syntactically well-formed, (2) it belongs to the scientific lexicon (3) it belongs to the domain lexicon (here linguistics), (4) it is a TO. Thus, TOs have been validated at each step.

For evaluation purposes, the dataset has been split into eight folds to apply leave-one-out cross-validation. Each training sub-corpora contains 48 documents and the corresponding evaluation sub-corpora contains 7 documents. As far as possible, the six subdomains of the corpus (Language Acquisition, Lexicon, Descriptive Linguistics, Linguistics and Language diseases, NLP and Sociolinguistics) are equi-distributed in the eight folds.

4. Methods

In this section, we first introduce a baseline and then present three original methods for term ambiguity diagnosis.

4.1. Baseline

This baseline is a simplified version of the Lesk Algorithm (Lesk, 1986). The class for a given TC occurrence is obtained by comparing its neighbourhood to the neighbourhood of its TOs and $NTOs$ in the training corpus. The neighborhood of an occurrence is the set of words occurring in the same XMLblock (paragraph, title...). Let N_{cand} be the neighborhood of a candidate in the test corpus. In the same fashion, let N_{term} be the neighborhood of TOs (resp. $N_{nonterm}$ for $NTOs$). The intersection between N_{cand} and N_{term} is compared to the intersection between N_{cand} $N_{nonterm}$. The largest intersection gives the class for the occurrence. If intersections have the same size (for instance if a TC is not present in the training corpus) this is a case of indecisiveness, it is resolved as follows:

- Precision-Oriented Lesk (POL): indecisiveness cases are classified as $NTOs$ in order to favour precision;
- Recall-Oriented Lesk (ROL): indecisiveness cases are classified as TOs in order to favour recall.

4.2. Hypotheses Based Approach (HB)

This approach assumes that words and word annotations (POS tags...) surrounding a TC occurrence define a useful context for classifying it as a TO or a NTO. The main difference with Lesk is that neighbourhood for TOs are restricted to words that occurs only with TOs and not with $NTOs$ and *vice-versa*. Hypotheses (Kuznetsov, 2004; Kuznetsov, 2001) are linked to Formal Concept Analysis (FCA) and result from a symbolic machine learning approach based on itemset mining and a classification of positive (TOs) and negative ($NTOs$) examples. FCA is a data analysis theory which builds conceptual structures defined by means of the attributes shared by objects. Formally, this theory is based on the triple $K = (G, M, I)$ called *formal context*, where *G* is a set of objects, *M* is a set of attributes and *I* is the the binary relation $I \subseteq G \times M$ between objects and attributes. Therefore, *(g, m)* $\in$ *I* means that *g* has the attribute *m*. For instance, occurrences of the introductory examples with the TC `Aspect` are encoded in the formal context given by Table 1. A more detailed and formal description of the method and its results for ambiguity diagnosis are given in (Melo-Mora and Toussaint, 2015).

	is	a	category	which	express	the	internal	sequence	of	process	this	last	primordial	T_+	T_-
***Aspect*(S_1)**	x	x	x	x	x	x	x	x	x	x				x	
***Aspect*(S_2)**	x										x	x	x		x

Table 1: An example of formal context where each row represents one occurrence of the TC `Aspect` with the words appearing in its neighbourhood

Hypotheses are computed for each TC separately. For each TC, a set of positive and a set of negative hypotheses are built. First, each TO and NTO is described by its textual context, *i.e.* the words in the sentence. The occurrence is a "positive example" (belonging to the "T_+ class") if it is a TO or a "negative example" ("T_- class") if it is a NTO. A positive (resp. negative) hypothesis is an itemset of words corresponding to positive (resp. negative) occurrences of a TC.

With regard to FCA theory, this classification method can be described by three sub-contexts : a positive context $K_+ = (G_+, M, I_+)$, a negative context $K_- = (G_-, M, I_-)$, and an undetermined context $K_\tau = (G_\tau, M, I_\tau)$ that contains instances to be classified. M is a set of attributes (surrounding words), *T* is the target attribute and $T \notin M$, G_+ is the set of positive examples whereas G_- is the set of negative examples. Alternatively, G_τ denotes the set of new examples to be classified.

A positive hypothesis H_+ for T is defined as a non empty set of attributes of K_+ which is not contained in the description of any negative example $g \in G_-$. A *negative hypothesis* H_-, is defined accordingly. A positive hypothesis H_+ generalizes G_+ subsets and defines a cause of the target attribute T. In the best case, the membership to G_+ supposes a particular attribute combination (one hypothesis). However, in most cases it is necessary to find several attribute combinations *i.e.* several positive hypotheses to characterize G_+ examples. Ideally, we would like to find enough positive hypotheses to cover all G_+ examples. To reduce the number of hypotheses and in accordance with FCA, an hypothesis is a closed itemset: it corresponds to the maximal set of words shared by a maximal set of occurrences.

Thereby, hypotheses can be used to classify an undetermined example. If the description of x (*i.e.* words in the same sentence as x) contains at lest one positive hypothesis and no negative hypothesis, then, x is classified as a positive example. If the intent of x contains at least one negative

hypothesis and no positive hypothesis, then it is a negative example. Otherwise, x remains unclassified. It should be mentioned that several alternative strategies could manage these unclassified examples such as assigning an arbitrary positive or negative class. It could improve precision or recall but would contribute to confusing the analysis of the results.
In addition, we can restrict the number of useful hypotheses with regard to subsumption in the lattice. Formally, a positive hypothesis H_+ is a *minimal positive hypothesis* if there is no positive hypothesis H such that $H \subset H_+$. *Minimal negative hypothesis* is defined similarly. Hypotheses which are not minimal should not be considered for classification because they do not improve discrimination between positive and negative examples.

4.3. Lafon's Specificity Approach (LS)

This approach relies on a statistical analysis following Lafon's model of specificity (Lafon, 1980; Drouin, 2007). Two sets of lexical components are extracted from the training corpus: for each TC, a set of lexical contexts for its TO, the other one for its NTO.
The terminological set and the non-terminological set are built as follows:

- For each TC occurrence in the training corpus, if it is a TO (resp. NTO), store the lexical components of its linguistic context (the paragraph as it is marked by the well-known XML tag <p>) to the terminological (resp. non terminological) set;

- In each set, for each lexical unit, compute the specificity score. It reflects the over-representation or the under-representation of the unit inside each set in comparison with the whole corpus. This score is computed with the TXM tool (Heiden et al., 2010);

- Finally, each set contains pairs (lexical unit, specificity score). In Table 2, some of the most specific components of the terminological (resp. non-terminological) sets which are computed for the TC `aspect` on the training corpus are reproduced here.

TO pairs		NTO pairs	
England	41,16	*orientation*	19,55
past	34,21	*community*	11,45
English	28,73	*representation*	11,41
preterit	19,35	*competence*	10,34
future	17,40	*speaking*	8,91
achieved	16,61	*familiar*	8,84
duration	15,78	*spirit*	8,42
language	14,38	*plaful*	7,86
rule	11,10	*thing*	7,77
narration	10,87	*feature*	7,21
...		...	

Table 2: Lafon's Specificity: most specific components of the terminological (resp. non-terminological) sets for the TC `aspect`

The terminological pairs may lead to the conclusion that most papers in which `aspect` occurs with its linguistic meaning (the TO) are dealing with applied linguistics for non native speakers. By contrast, the diversity of the non-terminological pairs may only lead to the conclusion that papers in which `aspect` occurs with one of its non-terminological meanings, for instance a specific facet for a given issue, are dealing with many other issues.
But these sets are not intended to provide a meaningful representation of TO (resp. NTO) of the TC `aspect`. There are only used to decide, for each TC occurrence, if it is closer to TO (resp. NTO) of `aspect` following the sets which have been computed with the training corpus.
In the test corpus, the linguistic context of each TC occurrence is compared to these two sets.
The method selects the set with the most significant intersection: the largest number of units in common with the highest specificity score.
For instance, the following occurrence of `aspect`

> *L'***aspect**, *catégorie par laquelle l' énonciateur conçoit le déroulement interne d' un procès, est marqué en créole haïtien au moyen de particules marqueurs prédicatifs MP préposées au verbe.* ('*Aspect, category by which the speaker conveives the internal workflow of a process, is marked in Haitian Creole by means of particles predicative markers MP which precede the verb.*')

is considered as an TO because the intersection with the computed pairs on training corpus is more significant with TO pairs. Some of the most specific components which are shared are *present, workflow, to express, past, duration, language, to speak, etc.*.
By contrast, the following occurrence of `aspect`

> L'**aspect** différentiel cède la place à une vision positive substantielle du lexique. ('*The differential aspect give away to a positive substantial vision of the lexicon.*')

is considered as an NTO because the intersection with the computed pairs on training corpus is more significant with NTO pairs. Some of the most specific components which are shared are *orientation, relative, specific, spirit, community, unpublished, common, representation, etc.*.

4.4. Salience Approach (SA)

For term disambiguation, all generic machine learning classification algorithms are applicable: discriminative algorithms such as $C4.5$ (Quinlan, 1993) or aggregative algorithms such as Naives Bayes. In this approach, the features are the POS tag of the TC, its lemma and discourse clues that rely on text genre properties called **salience** ((Brixtel et al., 2013; Lejeune and Daille, 2015)). The assumption is that TO are more often used in salient positions.
Scientific texts contains only a few important terms. These terms appear in salient positions in order to ease the understanding of the reader. When an important term occurs it comes along with other important terms in a gregarious manner. On the contrary, $NTOs$ are more equally

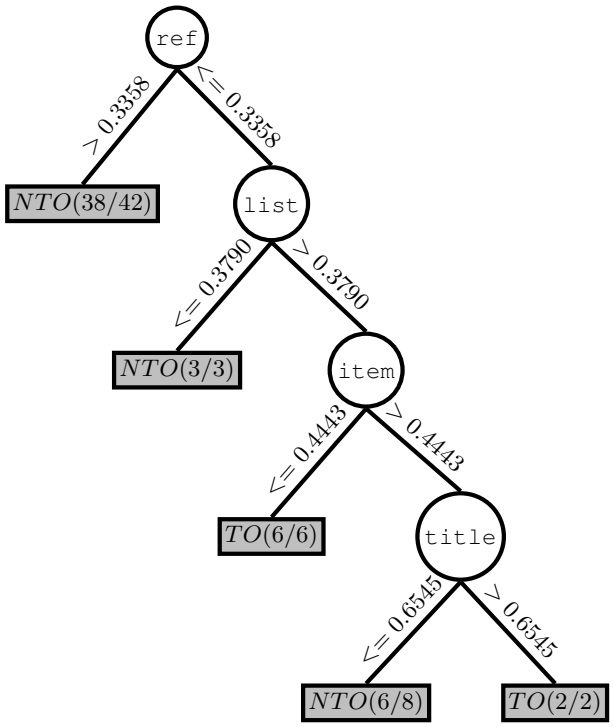

Figure 1: Decision Tree computed for occurrences of the TC `Aspect`: each node is an XML tag and each edge exhibits the normalized distance between on occurrence and the closest tag of this type, for each decision (TO or NTO), the proportion of True Positives is given

distributed within the document. Furthermore, the number of salient positions is limited so that it is unlikely that $NTOs$ will occupy salient positions rather than other positions. The discourse features are salient positions that are computed by taking advantage of the XML structure. The main tags found in our corpus are :

text the full text, including its **title** and its **body** ;

div a section with **head** its title and **p** its paragraphs ;

list a bulleted list with **item** its items;

keywords keywords given by authors ;

ref reference to bibliography.

An example of a decision tree obtained for the term `Aspect` is given in Figure 1. This is a set of rules specific to occurrences of this TC that are not classified using generic rules. This example shows for instance that occurrences of `Aspect` are very unlikely to be TOs when they are not close to bibliographical references represented by the ref tag (Node 1).
For each TC, the position is computed as follows:

- For each XML tag type in the document:
  - Compute the distance (in characters) between the TC and the closest tag of this type;
  - Normalize this distance with respect to the length of the text.

	POL	ROL	HB	LS	SA
DR	78.8%	**100%**	53.5%	71.8%	**100%**
P	69.8%	66.2%	**84.9%**	69.1%	73.0%
FN_A	5398	9841	**2374**	4996	6634
R_A	59.8%	53.1%	**78.9%**	66.9%	68.4%
F_{1_A}	64.4%	59.0%	**81.8%**	67.9%	70.6%
$F_{0.5_A}$	65.4%	60.3%	**82.48%**	68.2%	71.1%
FN_B	12955	9841	12125	10914	**6634**
R_B	38.3%	53.1%	42.2%	48.0%	**68.4%**
F_{1_B}	49.4%	59.0%	56.4%	56.6%	**70.6%**
$F_{0.5_B}$	52.5%	60.3%	60.56%	58.8%	**71.1%**

Table 3: Results for the two baselines, Precision Oriented (POL) and Recall Oriented (ROL) Lesk, and the three approaches: Hypotheses Based (HB), Lafon's Specificity (LS) and Salience (SA) for the two settings

Positional features are combined with lemmas and POS tags to train a classifier. For the choice of a classifier, we rely on the work of (Yarowsky and Florian, 2002) that observed that discriminative algorithms such as decision trees perform better than aggregative algorithms for smaller sets of highly discriminative features, and use the default settings of $C4.5$ included in the WEKA tool (Witten and Fanck, 2005).

5. Results

The results obtained on the test corpus are presented in Table 3. True Positives (TPs) are correctly classified TOs. False Positives (FPs) are $NTOs$ wrongly tagged as TOs. Some of the methods (POL, LS and HB) do not give an answer for every candidate, this indecisiveness leads to unclassified TOs. This may affect computation and analysis of the recall scores. Therefore, we propose two definitions of False Negatives (FN).
Type A False Negatives (FN_A) are misclassified TOs only, they are used for the A setting. By adding unclassified TOs and misclassified TOs, we obtain type B False Negatives (FN_B), used for the B setting.
The A setting favours precision-oriented approaches that do not decide for every occurrence. On the contrary, the B setting favours recall-oriented approaches. We also computed the decision rate which is the number of TOs for which a decision is taken.
The measures are computed as follows:

- Decision Rate: $DR = (TP + FN_A)/(TP + FN_B)$
- Precision : $P = TP/(TP + FP)$
- Type A Recall : $R_A = TP/(TP + FN_A)$
- Type B Recall : $R_B = TP/(TP + FN_B)$
- F_A-measure: $F_{\beta_A} = (1 + \beta^2) * \frac{P_A * R_A}{(\beta^2 * P_A) + R_A}$
- F_B-measure: $F_{\beta_B} = (1 + \beta^2) * \frac{P_B * R_B}{(\beta^2 * P_B) + R_B}$

F-measure is computed with the classical setting $\beta = 1$ and with $\beta = 0.5$ to give a greater importance to precision.

6. Discussion and Conclusion

In this section we first present an analysis of the behavior of HB and SA methods. We then compare the results given by the three methods relatively to manual annotation (MA).

Analysing Hypotheses The number of positive hypotheses and negative hypotheses varies a lot depending on the TC. Table 4 gives observations for five TCs. *Frequency* is the number of occurrences in the training set, among them some are *positive occurrences* and the ratio between these two values gives the *terminological degree*. The tables gives also the total number of surrounding words involved in positive (resp. negative) hypotheses and the number of positive (resp. negative) hypotheses. Unsurprisingly, the number of hypotheses mainly increases (even if monotonicity cannot be ensured) with the number of examples. The number of hypotheses is usually much higher than the number of examples. The number of positive hypotheses varies in a non-monotonic way respectively to the terminological degree. However, if a term has a high terminological degree, the number of positive hypotheses is greater that the number of negative hypotheses and reciprocally for low terminological degree. When the terminological degree is around 50%, positive/negative numbers of hypotheses are rather balanced .

TC	Frequency	Positive Occurrences	Terminolog. Degree	Words used only in T_+	Hypotheses Generated from T_+	Proportion of Positive Hypotheses	Shared Words	Negative Examples	Words used only in T_-	Hypotheses Generated from T_-
adjective	216	207	95.83%	966	301	97.41%	64	9	59	8
corpus	688	510	74.12%	1035	1347	81.93%	713	178	535	297
text	568	266	46.83%	735	913	52.32%	772	302	792	832
relation	676	171	25.29%	159	183	11.48%	629	505	1427	1410
semantic	413	80	19.37%	272	108	8.88%	560	333	1258	1107

Table 4: Classification summary of candidate occurrences

Results from the test phase for hypotheses are given in Table 5 for some TCs. Table 5 shows the average number of hypotheses used to tags TC occurrences over the different runs where `Ex2tag` is the average number of occurrences to be tagged for this TC.

As hypotheses are set of surrounding words, we can observe what are the positive and the negative triggers for a given TC. Table 6 gives some examples of hypotheses for the TC `argument`. They have been ranked in decreasing order of the *stability* measure, a measure defined in FCA for evaluating quality of concepts in a lattice (Kuznetsov, 2007).

Analysis of the Salience Approach The POS tag feature is usually very useful when a TC can be an adjective or a noun: when a TC is an adjective, it is more likely to be a NTO. For instance, `radical` as an adjective is a NTO in 90% of times while it is a TO in 80% of times when it is a noun. Another interesting example is the TC `linguistique` (linguistics), its occurrences in the beginning of a paragraphs are always TOs. The classification of its other occurrences implies rules combining the POS tag feature and the salience features.

TC	Term. Degree	Ex2tag	Hypotheses used positive	Hypotheses used negative	Unclassified examples positive	Unclassified examples negative
adjective	95.83%	27	46	0	1.375	0.125
corpus	74.12%	86	268	61	25.5	5.25
text	46.83%	71	194	145	7.625	5.75
relation	25.29%	142.25	16	244	1.5	9.5
semantic	19.37%	51.62	20	288	0.5	5.5

Table 5: Average for all the runs of hypotheses used in test and unnamed examples in k-fold cross-validation ($k = 8$)

Support	Stability	Hypotheses in T_+	Hypotheses in T_+ *-english-*
Positive Hypotheses			
7	0.7968	[sdrt, être, argument]	[sdrt, be, argument]
9	0.7792	[argument, plus]	[argument, more]
6	0.73437	[être, argument, aussi]	[be, argument, also]
6	0.7187	[argument, verbal]	[argument, verbal]
...	...	...	...
5	0.6562	[être, argument, indique]	[be, argument, denote]
4	0.5	[argument, syntaxique]	[argument, syntactic]
...	...	...	...
6	0.3281	[être, argument, rst]	[be, argument, rst]
8	0.25	[argument, nucleus]	[argument, nucleus]
Negative Hypotheses			
3	0.5	[argument, prendre]	[argument, assume]
1	0.5	[trancher, pas, ne, argument, permettre, décisif, position, avoir]	[settle, not, argument, allow, decisive, position, have]
...	...	...	...
4	0.375	[argument, hypothèse]	[argument, hypothesis]
4	0.3125	[dire, argument]	[say, argument]
...	...	...	...
2	0.25	[trouver, même, argument]	[find, same, argument]

Table 6: Some positive and negative hypotheses for the `argument` candidate term

In order to ease comparisons with the hypothesis-based method, Table 7 exhibits some results for the TC `semantic` already analyzed in Table 5. Less than 20% of its occurrences are TOs, they are equally distributed among the two POS tags but 77% of TOs as a noun whereas it is the case only for 15% of its occurrences as an adjective. These examples were found in weakly structured documents with few references and very long sections. These are critical cases for our method when neither the POS tag nor structural features give enough information for the classification. The method still gives a diagnosis but is not reliable.

Analysis of the Results We can see an important difference when examining the performances in the two settings. In the A setting, the HB approach gives the best results thanks to its greater precision. The two baselines are outperformed: their type A recall is low and precision is outperformed by both SA and HB approaches. Conversely,

Res	POS	title	head	p	item	ref
FN	ADJ	0.8131	0.5064	0.1914	0.8079	0.4875
FN	NOUN	0.8324	0.5439	0.1116	0.8279	0.1516
TN	ADJ	0.8368	0.5483	0.1160	0.8323	0.1472
TP	NOUN	0.8406	0.5717	0.1636	0.8318	0.0342
TN	ADJ	0.8523	0.5638	0.1315	0.8478	0.1317

Table 7: Examples of good and bad classification of the TC `semantic`

Confidence (in %)	Support (in %)	Association rule
0.93	0.31	[HB-NTO, LS-NTO] ⟶ [SA-NTO]
0.93	0.28	[MA-NTO, HB-NTO, LS-NTO] ⟶ [SA-NTO]
0.93	0.1	**[HB-TO, LS-TO] ⟶ [SA-TO]**
0.93	0.09	[MA-TO, HB-TO, LS-TO] ⟶ [SA-TO]
0.92	0.33	[MA-NTO, HB-NTO] ⟶ [SA-NTO]
0.92	0.13	[MA-TO, HB-TO] ⟶ [SA-TO]
0.91	0.4	**[MA-NTO, LS-NTO] ⟶ [SA-NTO]**
0.91	0.36	**[HB-NTO] ⟶ [SA-NTO]**
0.91	0.28	**[SA-NTO, HB-NTO, LS-NTO] ⟶ [MA-NTO]**
0.91	0.16	**[HB-TO] ⟶ [SA-TO]**
0.91	0.09	[MA-NTO, HB-UN, LS-UN] ⟶ [SA-NTO]
0.9	0.36	[HB-NTO] ⟶ [MA-NTO]
0.9	0.33	[SA-NTO, HB-NTO] ⟶ [MA-NTO]
0.9	0.3	[HB-NTO, LS-NTO] ⟶ [MA-NTO]
0.9	0.11	[MA-NTO, HB-UN, LS-NTO] ⟶ [SA-NTO]
0.89	0.15	[MA-TO, LS-TO] ⟶ [SA-TO]
0.89	0.11	[MA-NTO, LS-UN] ⟶ [SA-NTO]
0.88	0.4	[SA-NTO, LS-NTO] ⟶ [MA-NTO]
0.88	0.05	[MA-TO, HB-UN, LS-TO] ⟶ [SA-TO]
…	…	…
0.85	0.11	**[SA-NTO, HB-UN, LS-NTO] ⟶ [MA-NTO]**
0.85	0.09	**[SA-TO, HB-TO, LS-TO] ⟶ [MA-TO]**
…	…	…
0.7	0.09	**[SA-NTO, HB-UN, LS-UN] ⟶ [MA-NTO]**
…	…	…

Table 8: Some best-confidence association rules between annotations

the B setting favours methods with a higher decision rate. It should be noticed that no strategy has been implemented yet to compute a value when our algorithms (LS and HB) are in a situation of indecisiveness while our Lesk version uses one.

We performed a more fine-grained comparison of the results of the approaches, occurrence per occurrence. Each occurrence of a TC, whatever the TC is, is described by a set of four multi-valued attributes (the four different annotations) corresponding the manual annotations (MA), the hypotheses-based annotations (HB), Lafon's specificity (LS) and Salience (SA) annotations. Among 59,168 occurrences, 24,964 are not classified by HB, and 13,615 are not classified by LS. 10,230 are not classified by these two approaches. Among these 10,230 occurrences, 1,988 are TO correctly classified by SA, 5,549 are NTO correctly classified by SA, and 2,287 are TO wrongly classified by SA.

To study links between the different methods, we also extracted association rules between the diagnosis given by the different methods. Each occurrence of TC has been described following this example :

#d1e2267-definition : MA-NTO, SA-NTO, HB-UN, Laf-UN

This can be read as follows:

- the identifier of the occurrence: this is the occurrence with the ID `#d1e2267` of the TC `definition`;
- the manual annotation (MA), salience annotation (SA), hypotheses annotation (HB) and Lafon's specificity annotation (LS)
- and the value of the annotation: non-terminological (`NTO`), terminological(`TO`) or unknown (`UN`)

An association rule of the type :

`confidence` $=0.93$, `support` $=0.1$,
[HB-TO, LS-TO] ⟶ [SA-TO]

means that if HB and LS annotates an occurrence as a TO then SA will do so in 93% of the cases (`confidence`) and it concerns 5916 occurrences, *i.e.* 10% (`support`) of the total number of the occurrences (59,168).

We extracted 86 association rules with a confidence higher than 50% and a support higher than 5% (2956 occurrences). Table 6. shows some of the best-confidence rules. One should be aware that association rules do not express causality but only observations between annotations. Among these association rules, let us give a focus on :

- 0.91, 0.4, [MA-NTO, LS-NTO] ⟶ [SA-NTO] means that when MA and LS annotates an occurrence as NTO, then SA mostly does so.
- 0.91, 0.36, [HB-NTO] ⟶ [SA-NTO] means that if HB gives a NTO diagnosis then SA mostly does so.
- 0.91, 0.28, [SA-NTO, HB-NTO, LS-NTO] ⟶ [MA-NTO] means that if the three methods agree on a NTO diagnosis, then generally MA is NTO.
- 0.91, 0.16, [HB-TO] ⟶ [SA-TO] means that if HB gives a TO diagnosis, then SA mostly does so.
- 0.85, 0.11, [SA-NTO, HB-UN, LS-NTO] ⟶ [MA-NTO] means that if SA and SL agree on a NTO diagnosis, then probably MA is NTO.
- 0.85, 0.09, [SA-TO, HB-TO, LS-TO] ⟶ [MA-TO] means that if our three methods agree on a TO annotation for an occurrence, then generally MA is TO.
- 0.7, 0.09, [SA-NTO, HB-UN, LS-UN] ⟶ [MA-NTO] means that when SA gives a NTO diagnosis when HB and LS cannot take decision, then it is not always a good diagnosis (confidence is only 0.7).

Conclusion In this paper, we presented a dataset designed for an ambiguity diagnosis task. We evaluated three methods and two baselines derived from the Lesk algorithm. We pointed out that a difficulty for evaluating this task is the impact of two different types of FNs: misclassified items VS unclassified items. We showed that this has a great impact on evaluation.

In future work, we will combine the different methods in order to take advantage of their different properties in terms of confidence (precision) and coverage (recall). We observed that a combination of our three methods of annotation that roughly favours TO annotations will pull down precision very close to the worst precision of the three methods and will provide a very low improvement of recall. Thus, association rules could probably suggest a better combination of these three methods.

7. References

Brixtel, R., Lejeune, G., Doucet, A., and Lucas, N. (2013). Any Language Early Detection of Epidemic Diseases from Web News Streams. In *International Conference on Healthcare Informatics (ICHI)*, pages 159–168.

Cothire-Robert, D. (2007). Stratégies des restitutions des constructions verbales sérielles du créole hatien en franais l2. In *Autour des langues et du langage: perspective pluridisciplinaire*. Presses Universitaires de Grenoble.

Daille, B., Jacquin, C., Monceaux, L., Morin, E., and Rocheteau, J. (2011). TTC TermSuite : une chaîne de traitement pour la fouille terminologique multilingue. In *18ème Conférence francophone sur le Traitement Automatique des Langues Naturelles Conference (TALN 2011).*, Montpellier, France, June. Démonstration.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 12(2):45–64.

El-Khoury, T. (2007). Les procédés de métaphorisation dans le discours médical arabe : étude de cas. In *Autour des langues et du langage: perspective pluridisciplinaire*. Presses Universitaires de Grenoble.

Foo, J. and Merkel, M. (2010). Using machine learning to perform automatic term recognition. In Núria Bel, et al., editors, *LREC 2010 Workshop Methods for the automatic acquisition of Language Resources and their evaluation methods*, pages 49–54, Malta.

Gaiffe, B., Husson, B., Jacquey, E., and Kister, L. (2015). Smarties: Consultation des fichiers annotés manuellement, domain scientext 2014, available at http://apps.atilf.fr/smarties/index.php?r=text/listtext. Technical report.

Heiden, S., Magué, J.-P., and Pincemin, B. (2010). Txm : Une plateforme logicielle open-source pour la textométrie conception et développement. In *Proceedings of JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, page 12pp, Rome, Italie.

Judea, A., Schütze, H., and Bruegmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 290–300, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Kuznetsov, S. O. (2001). Machine learning on the basis of formal concept analysis. *Autom. Remote Control*, 62(10):1543–1564.

Kuznetsov, S. O. (2004). Complexity of learning in concept lattices from positive and negative examples. *Discrete Applied Mathematics*, 142(13):111 – 125. Boolean and Pseudo-Boolean Functions.

Kuznetsov, S. O. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49(1-4):101–115.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1:127–165.

Lejeune, G. and Daille, B. (2015). Vers un diagnostic d'ambiguïté des termes candidats d'un texte. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*, pages 446–452.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

L'Homme, M.-C. (2004). *La terminologie : principes et techniques*. Presses de l'Universit de Montral.

Melo-Mora, L.-F. and Toussaint, Y. (2015). Automatic validation of terminology by means of formal concept analysis. In *International Conference on Formal Concept Analysis (ICFCA)*.

Navigli, R. (2009). Word sens disambiguation: A survey. *ACM Computing Surveys*, 41(2).

Quinlan, J. R. (1993). *C4.5: programms for machine learning*. Morgan Kaufmann Publishers, San Francisco, CA, USA.

Tutin, A. and Grossmann, F. (2015). Scientext: Un corpus et des outils pour étudier le positionnement et le raisonnement dans les écrits scientifiques, available at `http://scientext.msh-alpes.fr/scientext-site/spip.php?article8`.

Witten, I. H. and Fanck, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Francisco.

Yarowsky, D. and Florian, R. (2002). Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310.