Identification of Ambiguous Multiword Expressions Using Sequence Models and Lexical Resources

Manon Scholivet and **Carlos Ramisch**
Aix Marseille Univ, CNRS, LIF, Marseille, France
manon.scholivet@etu.univ-amu.fr
carlos.ramisch@lif.univ-mrs.fr

Abstract

We present a simple and efficient tagger capable of identifying highly ambiguous multiword expressions (MWEs) in French texts. It is based on conditional random fields (CRF), using local context information as features. We show that this approach can obtain results that, in some cases, approach more sophisticated parser-based MWE identification methods without requiring syntactic trees from a treebank. Moreover, we study how well the CRF can take into account external information coming from a lexicon.

1 Introduction

Identifying multiword expressions (MWEs) in running text with the help of a lexicon is often considered as a trivial task. In theory, one could simply scan the text once and mark (e.g. join with an underscore) all sequences of tokens that appear in the MWE lexicon. Direct matching and projection of lexical entries onto the corpus can be employed as a preprocessing step in parsing and MT (Nivre and Nilsson, 2004; Carpuat and Diab, 2010). Afterward, MWEs can be retokenized and treated as words with spaces, improving parsing and MT quality.

However, this simple pipeline does not work for many categories of MWEs, since variability and inflection may pose problems. For instance, if a lexicon contains the idiom *to make a face*, string matching will fail to identify it in *children are always making faces*. Since lexicons contain canonical (lemmatized) forms, matching must take inflection into account. This can be carried out by (a) pre-analysing the text and matching lemmas and POS tags instead of word forms (Finlayson and Kulkarni, 2011) or (b) using lexicons of inflected MWEs (Silberztein et al., 2012).

Things get more complicated when the target MWEs are ambiguous, though. An MWE is *ambiguous* when its member words can cooccur without forming an expression. For instance, *to make a face* is an idiom meaning ‘to show a funny facial expression’, but it can also be used literally when someone is making a snowman (Fazly et al., 2009). Additionally, the words of the expression can cooccur by chance, not forming a phrase (Boukobza and Rappoport, 2009; Shigeto et al., 2013). For example, *up to* is an MWE in *they accepted up to 100 candidates* but not in *you should look it up to avoid making typos*.

This paper focuses on a specific category of highly frequent and ambiguous MWEs in French. Indeed, in French some of the most recurrent function words are ambiguous MWEs. For instance, some conjunctions are formed by combining adverbs like *ainsi* (*likewise*) and *maintenant* (*now*) with subordinate conjunctions like *que* (*that*). However, they may also cooccur by chance when the adverb modifies a verb followed by a subordinate clause, as in the example taken from Nasr et al. (2015) :

1. *Je mange* **bien que** *je n’aie pas faim*
   I eat **although** *I am not hungry*
2. *Je pense* **bien** **que** *je n’ai pas faim*
   I think **indeed that** *I am not hungry*

The same happens for determiners like *de la* (partitive *some*), which coincides with preposition *de* (*of*) and determiner *la* (*the*).

3. *Il boit* **de la** *bière*
   He drinks **some** *beer*
4. *Il parle* **de** **la** *bière*
   He talks **about the** *beer*

As showed by Nasr et al. (2015), recognizing

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 167–175,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

these MWEs automatically requires quite high-level syntactic information such as access to a verbal subcategorization lexicon. Our hypothesis is that this information can be modeled without the use of a parser by choosing an appropriate data encoding and representative features.

The main reason why we are interested in these particular constructions is that they are frequent: in the frWaC corpus, containing 1.6 billion words, 2.1% of the sentences contain at least one occurrence of adverb+*que* construction, and 48.6% contain at least one occurrence of *de*+determiner construction. For example, the word *des* is the 7^{th} most frequent word in this corpus. Even if some of these constructions (*bien que, ainsi que*) are more frequent in formal registers, all the others are really pervasive and register-independent.

We propose a simple, fast and generic sequence model for tagging ambiguous MWEs using a CRF. One of the main advantages of the CRF is that we do not need a syntactic tree to train our model, unlike methods based on a parser. Moreover, for expressions that are not very syntactically flexible, it is natural to ask ourself if we really need a parser for this task. Parsers are good for discontiguous MWEs, but contiguous ones in theory can be modelled by sequence models that take ambiguity into account (such as CRFs). Regardless of the syntactic nature of these ambiguities, we expect that the CRF's highly lexicalised model compensates for the lack of structure. We focus on grammatical MWEs in French, which are prototypical examples of ambiguous MWEs. Our CRF-based approach pre-identifies MWEs without resorting to syntactic trees, and results are close to those obtained by state-of-the-art parsers (Green et al., 2013; Nasr et al., 2015). We also study the influence of features derived from an external lexicon of verb valence. We believe that our approach can be useful (a) when no treebank is available to perform parsing-based MWE identification and (b) as a preprocessing step to parsing which can improve parsing quality by reducing attachment ambiguities (Nivre and Nilsson, 2004).

2 Related Work

Token identification of ambiguous MWEs in running text can be modelled as a machine learning problem that learns from MWE-annotated corpora and treebanks. To date, it has been carried out using mainly two types of models: sequence taggers and parsers. Sequence taggers, like conditional random fields (CRFs), structured support vector machines and structured perceptron, allow disambiguating MWEs using local feature sets such as word affixes and surrounding word and POS n-grams. Parsers, on the other hand, can take longer-distance relations and features into account when building a parse tree, at the expense of using more complex models.

Sequence taggers have been proven useful in identifying MWEs. MWE identification is also sometimes included into part-of-speech (POS) taggers in the form of special tags. Experiments have shown the feasibility of sequence tagging for general expressions and named entities in English and Hungarian (Vincze et al., 2011), verb-noun idioms in English (Diab and Bhutada, 2009) and general expressions in French (Constant and Sigogne, 2011) and in English (Schneider et al., 2014). Shigeto et al. (2013) tackle specifically English function words and build a CRF from the Penn Treebank, additionally correcting incoherent annotations. We develop a similar system for French, using the MWE annotation of the French Treebank as training data.

Parsing-based MWE identification requires a treebank annotated with MWEs. Lexicalized constituency parsers model MWEs as special non-terminal nodes included in regular rules (Green et al., 2013). In constituency parsers, it is possible to employ a similar approach, using special dependency labels to identify relations between words that make up an expression (Candito and Constant, 2014). This technique has shown good performance in identifying ambiguous grammatical MWEs in French (Nasr et al., 2015).

Our paper adapts a standard CRF model like the ones proposed by Constant and Sigogne (2011) and Shigeto et al. (2013) to deal with ambiguous contiguous MWEs. Our hypothesis is that sophisticated techniques like the ones described by Green et al. (2013) and Nasr et al. (2015) are not required to obtain good performances on these expressions.

3 CRF-Based MWE Tagger

We trained a CRF tagger using CRFSuite[1] (Okazaki, 2007). We used a modified version of the French Treebank (Abeillé et al., 2003) as train-

[1] `http://www.chokkan.org/software/crfsuite/`

i:	-2	-1	0	1	2	3
w$_i$:	*Il*	*jette*	*de*	*la*	*nourriture*	*périmée*
	He	*discards*	*some*		*food*	*expired*
MWE:	O	O	B	I	O	O

Figure 1: Example of BIO tagging of a sentence containing a *de*+determiner MWE.

ing data and the MORPH dataset[2] (Nasr et al., 2015) as development and test data. We also include features from an external valence lexicon, Dicovalence[3] (van den Eynde and Mertens, 2003). Since our focus is on function words, our evaluation covers adverb+*que* and *de*+determiner constructions present in the MORPH dataset.

Training Corpus The training corpus is an adaptation of the French Treebank (FTB) in CONLL format that we have transformed into the CRFsuite format. For each word, the corpus contains its wordform, lemma, POS (15 different coarse POS tags), and syntactic dependencies (that were ignored). In the original corpus, MWE information is represented as words with spaces. We have added an extra column containing MWE annotation using a Begin-Inside-Outside (BIO) encoding, as in Figure 1.

The MWE-BIO tags were generated using the following transformation heuristics:

- For adverb+*que* pairs (AQ):
  1. We scan the corpus looking for the lemmas *ainsi_que*, *alors_que*, *autant_que*, *bien_que*, *encore_que*, *maintenant_que* and *tant_que*.
  2. We split them in two new words and tag the adverb as B and *que* as I.
- For *de*+determiner pairs (DD):
  1. We scan the corpus looking for the wordforms *des*, *du*, *de_la* and *de_l'*. Due to French morphology, *de* is sometimes contracted with the articles *les* (determinate plural) and *le* (determinate singular masculine). Contractions are mandatory for both partitive and preposition+determiner uses. Therefore, we systematically separate these pairs into two tokens.
  2. If a sequence was tagged as a determiner (D), we split the tokens and tag *de* as B and the determiner as I.
  3. Contractions (*des*, *du*) tagged as P+D (preposition+determiner) were split in two tokens, both tagged as O.
- All other tokens are tagged as O, including some other types of MWEs.

The expressions under study in this paper are strictly continuous. In unreported experiments, we use the method described in (Schneider et al., 2014) to treat discontinuous MWEs (more informations in Section 5).

For the newly created tokens, we assign individual lemmas and POS tags. The word *de* is systematically tagged as P (preposition), not distinguishing partitives from prepositions at the POS level. The input to the CRF is a file containing one word per line, BIO tags as targets, and `featureName=value` pairs including n-grams of wordforms, lemmas and POS tags.

Development and Test Corpora To create our test and development (dev) corpora, we used the MORPH dataset. It contains a set of 1,269 example sentences of 7 ambiguous adverb+*que* constructions and 4 ambiguous *de*+determiner constructions. For each target construction, around 100 sentences extracted from the frWaC corpus were manually annotated as to whether they contain a multiword function word (MORPH) or accidental cooccurrence (OTHER). We have preprocessed the raw sentences as follows:

1. We have automatically POS tagged and lemmatized all sentences using an off-the-shelf POS tagger and lemmatizer independently trained on the FTB.[4] This information is used as features for our CRF.

2. We have located the target construction in the sentence and added BIO tags according to the manual annotation provided: target pairs in

[2] http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/morph

[3] http://bach.arts.kuleuven.be/dicovalence/

[4] http://macaon.lif.univ-mrs.fr/

MORPH sentences were tagged B + I, target pairs in OTHER sentences were tagged O.

3. For each target construction, we have taken the first 25 sentences as development corpus (dev, 275 sentences).

4. We created four targeted datasets: DEV_{AQ}, DEV_{DD}, FULL_{AQ} and FULL_{DD}, where the different construction classes are separated, in order to perform feature selection.

External Lexicon The verbal valence dictionary Dicovalence specifies the allowed types of complements per verb sense in French. For each verb, we extract two binary flags:

- `queCompl`: one of the senses of the verb has one object that can be introduced by *que*.[5]

- `deCompl`: one of the senses of the verb has a locative, temporal or prepositional paradigm that can be introduced by *de*.[6]

CRF Features We selected 37 different features (referred to as `ALL`) inspired on those proposed by Constant and Sigogne (2011):

- Single-token features ($\texttt{t}_\texttt{i}$):[7]
  - $\texttt{w}_0$: wordform of the current token.
  - $\texttt{l}_0$: lemma of the current token.
  - $\texttt{p}_0$: POS tag of the current token.
  - $\texttt{w}_\texttt{i}$, $\texttt{l}_\texttt{i}$ and $\texttt{p}_\texttt{i}$: wordform, lemma or POS of previous ($\texttt{i} \in \{-1, -2\}$) or next ($\texttt{i} \in \{+1, +2\}$) tokens.

- N-gram features ($\texttt{t}_\texttt{i-1}\texttt{t}_\texttt{i}$ and $\texttt{t}_\texttt{i-1}\texttt{t}_\texttt{i}\texttt{t}_\texttt{i+1}$):
  - $\texttt{w}_\texttt{i-1}\texttt{w}_\texttt{i}$, $\texttt{l}_\texttt{i-1}\texttt{l}_\texttt{i}$, $\texttt{p}_\texttt{i-1}\texttt{p}_\texttt{i}$: wordform, lemma and POS bigrams of previous-current ($\texttt{i}= 0$) and current-next ($\texttt{i}= 1$) tokens.
  - $\texttt{w}_\texttt{i-1}\texttt{w}_\texttt{i}\texttt{w}_\texttt{i+1}$,$\texttt{l}_\texttt{i-1}\texttt{l}_\texttt{i}\texttt{l}_\texttt{i+1}$,$\texttt{p}_\texttt{i-1}\texttt{p}_\texttt{i}\texttt{p}_\texttt{i+1}$: wordform, lemma and POS trigrams of previous-previous-current ($\texttt{i}= -1$), previous-current-next ($\texttt{i}= 0$) and current-next-next ($\texttt{i}= 1$) tokens.

- Orthographic features (`orth`):
  - `hyphen` and `digits`: the current word contains a hyphen or digits.
  - `f-capital`: the first letter of the current word is uppercase.
  - `a-capital`: all letters of the current word are uppercase.
  - `b-capital`: the first letter of the current word is uppercase, and it is at the beginning of a sentence.

- Lexicon features/Subcat features (`SF`):[8]
  - `queV`: the current word is *que*, and the closest verb to the left accepts a `queCompl`.
  - `deV`: the current word is *de*, and the closest verb to the left accepts a `deCompl`.

In our evaluation, we report precision (P), recall (R) and F-measure (F_1) of MWE tags. In other words, instead of calculating a micro-averaged scores over all BIO tags, we only look at the proportion of correctly guessed B tags. Since all our target expressions are composed of exactly 2 contiguous words, we can use this simplified score because all B tags are necessarily followed by exactly 1 I tag. As a consequence, the measured precision, recall and F-measure scores on B and I tags are identical.

4 Evaluation

We evaluate our approach in two experimental setups. First, we perform feature selection using the dev/test split of the MORPH dataset, both regarding coarse groups (4.1) and individual features (4.2). Then, we apply the best configuration to the whole MORPH dataset in order to compare our results with the state of the art (4.3).

4.1 Feature Selection: Coarse

Our first evaluation was performed on the dev sets for adverb+*que* (DEV_{AQ}, 175 sentences) and *de*+determiner (DEV_{DD}, 100 sentences). It includes all features described in Section 3 (`ALL`), and obtains an F_1 score of 75.47 for AQ and 69.7 for DD constructions, as shown in the first row of Table 1. The following rows of this table show the results of a first ablation study, conducted to identify coarse groups of features that are not discriminant and hurt performance.

[5] In Dicovalence, an object `P1, P2` or `P3` licenses a complementizer `qpind`

[6] In Dicovalence, the paradigm is `PDL`, `PT` or `PP`.

[7] $\texttt{t}_\texttt{i}$ is a shortcut denoting the group of features $\texttt{w}_\texttt{i}$, $\texttt{l}_\texttt{i}$ and $\texttt{p}_\texttt{i}$ for a token. The same applies to n-grams.

[8] This is the same as the *subcat feature* proposed by Nasr et al. (2015).

	DEV$_{AQ}$			DEV$_{DD}$		
Feature set	P	R	F_1	P	R	F_1
ALL	89.55	65.22	75.47	92.00	56.10	69.70
ALL − orth	90.28	70.65	79.27	95.83	56.10	70.77
ALL − W	90.79	75.00	82.14	87.10	65.85	75.00
ALL − SF	91.18	67.39	77.50	88.89	58.54	70.59
ALL − $\texttt{t}_{\pm 2}$	87.67	69.57	77.58	88.00	53.66	66.67
ALL − $\texttt{t}_{\texttt{i-1}}\texttt{t}_{\texttt{i}}\texttt{t}_{\texttt{i+1}}$	87.84	70.65	78.31	91.67	53.66	67.69
ALL − $\texttt{t}_{\texttt{i-1}}\texttt{t}_{\texttt{i}}$	**93.55**	63.04	75.32	95.83	56.10	70.77
ALL − $\texttt{t}_{\texttt{i-1}}\texttt{t}_{\texttt{i}}$ − $\texttt{t}_{\texttt{i-1}}\texttt{t}_{\texttt{i}}\texttt{t}_{\texttt{i+1}}$	88.57	67.39	76.54	**96.00**	58.54	72.73
ALL − orth − W	90.24	**80.43**	**85.06**	87.10	65.85	75.00
ALL − orth − W − $\texttt{t}_{\pm 2}$ (REF)	89.74	76.09	82.35	85.29	**70.73**	**77.33**

Table 1: First feature selection, removing coarse-grained feature groups.

	DEV$_{AQ}$			DEV$_{DD}$		
Features	P	R	F_1	P	R	F_1
REF	89.74	76.09	82.35	85.29	70.73	77.33
REF − SF	90.00	78.26	83.72	75.76	60.98	67.57
REF − $\texttt{t}_{\texttt{-1}}\texttt{t}_{\texttt{0}}$	90.54	72.83	80.72	85.29	70.73	77.33
REF − $\texttt{t}_{\texttt{0}}\texttt{t}_{\texttt{+1}}$	89.87	77.17	83.04	84.85	68.29	75.68
REF − $\texttt{t}_{\texttt{0}}\texttt{t}_{\texttt{+1}}\texttt{t}_{\texttt{+2}}$ (BEST)	87.36	82.61	84.92	83.78	75.61	79.49

Table 2: Second feature selection, removing fine-grained feature groups.

When we ignore orthographic features (ALL − orth), all scores increase for DEV$_{AQ}$ and DEV$_{DD}$, showing that MWE occurrences are not correlated with orthographic characteristics. F_1 also increases when we remove all wordform-level features, including single words and n-grams (represented by W). We hypothesize that the use of lemmas and POS is more adequate, since it reduces sparsity by conflating variants, so wordforms only introduce noise.

Then, we try to remove the subcat features (ALL − SF). This information seems important to us, because it allows assigning O tags to conjunctions and prepositions that introduce verbal complements. Surprisingly, though, the system performs better without them. We suppose that this happens because, since there are many features, the CRF disregards SF features anyway because they are not frequent enough. These features will be analyzed individually later (see Table 3).

Single tokens located 2 words apart from the target token should not provide much useful information, so we try to remove their corresponding features (ALL − $\texttt{t}_{\pm 2}$). While this is true for DEV$_{AQ}$, it does not hold for DEV$_{DD}$. Next, we try to remove all trigram, and then all bigram features at once. When we remove trigrams, F_1 decreases by 2.01 absolute points in DEV$_{DD}$ and increases by 2.84 absolute points in DEV$_{AQ}$. Bigrams are somehow included in trigrams, and their removal has little impact on the tagger's performance. When we remove bigram and trigram features altogether, scores are slightly better even though a large amount of information is ignored. Since these results are inconclusive, we perform a more fine-grained selection considering specific n-grams in Table 2.

Finally, we try to remove several groups of features at the same time. When we remove both orthographic and wordform features, F_1 increases to 85.06 for DEV$_{AQ}$ and 75.00 for DEV$_{DD}$. When we remove also tokens located far away from the current one, performance increases for DEV$_{DD}$ but not for DEV$_{AQ}$. Unreported experiments have shown, however, that further feature selection (Table 2) also has better results for DEV$_{AQ}$ when we ignore $\texttt{t}_{\pm 2}$ features. Therefore, our reference (REF) for the fine-grained feature selections experiments will be this set of features, corresponding to the last row of Table 1.

4.2 Feature Selection: Fine

In the second row of Table 2, we try to remove subcat features again from REF, because on Table 1 these features seem to hurt performance. However, this is not the case anymore. We assume that these features can be better taken into account now that there are less noisy features in the whole system.

The last three rows of the table show our experiments in trying to remove individual n-gram features that seemed not very informative or redundant to us. First, we delete the two types of bigram features independently, including wordforms, POS and lemmas. We can see that bigrams seem useful and their removal causes the scores to drop. The only exception are the results on DEV$_{AQ}$ for the bigram $\mathtt{t}_0\mathtt{t}_{+1}$.

Finally, we remove all trigram features of the form $\mathtt{t}_0\mathtt{t}_{+1}\mathtt{t}_{+2}$, . We can see that performance increases in both datasets. This makes sense because MWE identification generally does not depend on the next tokens, but on the previous ones. This is the best configuration obtained on the development datasets, and we will refer to it as BEST in the next experiments.

Our last feature selection experiments study the influence of subcategorization features individually, as shown in Table 3. We observe that deV is an important feature, because when we remove it, F_1 decreases by almost 7 absolute points on the DEV$_{DD}$ set. The feature queV, however, seems less important, and its absence only slightly decreases the F_1 score on the DEV$_{AQ}$ set. This is in line with what was observed by Nasr et al. (2015) for the whole dataset. In sum, these features seem to help but the system could benefit more from them with a more sophisticated representation.

Features	Dataset	P	R	F_1
BEST	DEV$_{AQ}$	87.36	82.61	84.92
	DEV$_{DD}$	83.78	75.61	79.49
BEST−queV	DEV$_{AQ}$	91.25	79.35	84.88
BEST−deV	DEV$_{DD}$	77.78	68.29	72.73

Table 3: Impact of subcat features (SF) on separate dev sets per construction.

4.3 Comparison with State of the Art

The best system obtained after feature selection was then compared with the results reported by Nasr et al. (2015) in Table 4. We include two versions of their systems since they also report experiments on including subcategorization features coming from Dicovalence.

We report the performance on the full MORPH dataset split in two parts: sentences containing adverb+*que* constructions (FULL$_{AQ}$) and sentences containing *de*+determiner constructions (FULL$_{DD}$). Even though the use of the full datasets is not ideal, given that we performed feature selection on part of these sentences, it allows direct comparison with related work.

We also report results of a simple baseline:

1. We extract from the French Treebank the list of all adverb+*que* and *de*+determiner pairs.

2. We calculate the proportion of times that they were annotated as MWEs (B-I tags) with respect to all their occurrences.

3. We keep in the list only those constructions annotated 50% of the time or more.

4. We systematically annotate these constructions as MWEs (B-I) in all sentences of the MORPH dataset, regardless of their context.

Table 4 shows that this baseline reaches 100% recall, covering all target constructions, but precision is very low due to the lack of context. Our BEST system can identify the target ambiguous MWEs much better than the baselines for both FULL$_{AQ}$ and FULL$_{DD}$.

We did not expect our system to outperform parsing-based approaches, which were trained on a full treebank, have access to more sophisticated models of a sentence's syntax, and handle long-distance relations and grammatical information. Nonetheless, for some constructions we obtain results that are near to those obtained by the parsers. For FULL$_{AQ}$, our BEST system obtains an F_1 score that is 1.2 absolute points lower than the best parser. For FULL$_{DD}$, however, our best system, which includes subcategorization features, is comparable with a parser without subcategorization features. When the parser has access to the lexicon, it beats our system a significant margin of 7.99 points, indicating that the accurate disambiguation of DD constructions indeed requires syntax-based methods rather than sequence taggers.

Despite the different performances depending on the nature of the target constructions, these results are encouraging, as they prove the feasibility of using sequence taggers for the identification of highly ambiguous MWEs. Our method has mainly

	FULL$_{AQ}$			FULL$_{DD}$		
System	P	R	F_1	P	R	F_1
Baseline	56.08	100.00	71.86	34.55	100.00	51.35
Nasr et al. (2015)−`SF`	88.71	82.03	85.24	77.00	73.09	75.00
Nasr et al. (2015)+`SF`	91.57	81.79	86.41	86.70	82.74	84.67
`BEST`	91.08	78.31	84.21	79.14	74.37	76.68

Table 4: Comparison with baseline and state of the art.

two advantages over parsing-based MWE identification: (a) it is fast and only requires a couple of minutes on a desktop computer to be trained and (b) it does not require the existence of a treebank annotated with MWEs.

Expression	P	R	F_1
ainsi que	94.44	93.15	93.79
alors que	84.00	97.67	90.32
autant que	93.48	51.81	66.67
bien que	100.00	91.43	95.52
encore que	76.19	94.12	84.21
maintenant que	97.62	64.06	77.36
tant que	100.00	60.00	75.00
de la	67.74	72.41	70.00
de les	92.41	71.57	80.66
de le	78.05	71.11	74.42
de l'	61.11	95.65	74.58

Table 5: Performance of the `BEST` configuration broken down by expression.

Table 5 shows the detailed scores for each expression in the MORPH dataset. We notice that some expressions seem to be particularly hard, specially if we look at precision, whereas for others we obtain performances well above 90%. When we compare our results to those reported by Nasr et al. (2015), we can see that they are similar to ours: *ainsi*, *alors* and *bien* have F_1 higher than 90%, while *autant* and *tant* are less than 80%. The adverb+*que* constrictions with *encore* and *maintenant* are the only ones which behave differently: our system is better for *encore*, but worse for *maintenant*. Likewise, for *de*+determiner expressions, our system obtains a performance that is near to their system without subcategorization features: both approaches are more efficient to identify the plural article *de les* than the partitive constructions.

5 Conclusions and Future Work

We have described and evaluated a simple and fast CRF tagger that is able to identify highly ambiguous multiword expressions in French[9]. We have reported a feature selection study and shown that, for adverb+*que* constructions, our results are near those obtained by parsers, even though we do not use syntactic trees. While these experiments shed some light on the nature of this frequent phenomenon in French, the methodology is highly empirical and cannot be easily adapted to other contexts. Therefore, we would like to experiment different techniques for generic automatic feature selection and classifier tuning (Ekbal and Saha, 2012). This could be performed on a small development set and ease the adaptation of the tagger to other contexts.

We also think it could be interesting to test more sophisticated baselines. For instance, we could learn simple conditional rules from the training corpus depending on the lemma of the preceding verb.

Another idea for future work is to study the interplay between automatic POS tagging and MWE identification. We recall that our results were obtained using an off-the-shelf POS tagger and lemmatizer. Potentially, performing both tasks jointly could help obtaining more precise results (Constant and Sigogne, 2011).

Moreover, we are not fully satisfied with the representation of subcategorization features. We would like to study why SF features are not very useful by looking at the verbs preceding the MWEs and their feature values, performing error analysis. Furthermore, we would like to try implementing a threshold on the distance between the verb and the MWE to tag: only verbs close enough to the target construction generate subcategorization features for the MWE candidate.

We would also like to perform a cross validation

[9] The system described in this paper is publicly available `http://mwetoolkit.sourceforge.net`

experience, training the system on the MORPH dataset itself instead of using the French Treebank. This would allow us to quantify to what extent the CRF is able to generalize from the training data, even if it has never seen a particular expression before but only similar ones.

Finally, we would also like to experiment with other sequence tagging models such as recurrent neural networks. In theory, such models are very efficient to perform feature selection and can also deal with continuous word representations, which can include semantic information. Moreover, distributed word representations are helpful in building cross-lingual MWE identification systems.

Acknowledgments

This work has been partly funded by projects PARSEME (Cost Action IC1207), PARSEME-FR (ANR-14-CERA-0001), and AIM-WEST (FAPERGS-INRIA 1706-2551/13-7).

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In Anne Abeillé, editor, *Treebanks: building and using parsed corpora*, pages 165–168. Kluwer academic publishers, Dordrecht, The Netherlands.

Ram Boukobza and Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 468–477, Singapore, August. Association for Computational Linguistics.

Marie Candito and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proc. of the 52nd ACL (Volume 1: Long Papers)*, pages 743–753, Baltimore, MD, USA, Jun. ACL.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California, Jun. ACL.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In Kordoni et al. (Kordoni et al., 2011), pages 49–56.

Mona Diab and Pravin Bhutada, 2009. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, chapter Verb Noun Construction MWE Token Classification, pages 17–22. Association for Computational Linguistics.

Asif Ekbal and Sriparna Saha. 2012. Multiobjective optimization for classifier ensemble and feature selection: an application to named entity recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(2):143–166.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Comp. Ling.*, 35(1):61–103.

Mark Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In Kordoni et al. (Kordoni et al., 2011), pages 20–24.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Comp. Ling.*, 39(1):195–227.

Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors. 2011. *Proc. of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA, Jun. ACL.

Alexis Nasr, Carlos Ramisch, José Deulofeu, and André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1116–1126. Association for Computational Linguistics.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *MEMURA 2004 – Methodologies and Evaluation of Multiword Units in Real-World Applications (LREC Workshop)*, pages 39–46.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Nathan Schneider, Emily Danchik, Chris Dyer, and A. Noah Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pages 193–206.

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the 9th Workshop on MWEs (MWE 2013)*, pages 139–144, Atlanta, GA, USA, Jun. ACL.

Max Silberztein, Tamás Váradi, and Marko Tadić. 2012. Open source multi-platform NooJ for NLP. In *Proc. of COLING 2012: Demonstration Papers*, pages 401–408, Mumbai, India, Dec. The Coling 2012 Organizing Committee.

Karel van den Eynde and Piet Mertens. 2003. La valence: l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, (13):63–104.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. Detecting noun compounds and light verb constructions: a contrastive study. In Kordoni et al. (Kordoni et al., 2011), pages 116–121.