CLASSiC

D2.3: Semantic Parser for French, built using one for English

Lonneke van der Plas, James Henderson, Paola Merlo

Distribution: Public

CLASSiC
Computational Learning in Adaptive Systems for Spoken Conversation
216594 Deliverable 2.3

October 2009

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development

The deliverable identification sheet is to be found on the reverse of this page.

Project ref. no.	216594
Project acronym	CLASSiC
Project full title	Computational Learning in Adaptive Systems for Spoken Conversation
Instrument	STREP
Thematic Priority	Cognitive Systems, Interaction, and Robotics
Start date / duration	01 March 2008 / 36 Months

Security	Public
Contractual date of delivery	M18 = Aug 2009
Actual date of delivery	October 2009
Deliverable number	2.3
Deliverable title	D2.3: Semantic Parser for French, built using one for English
Type	Prototype
Status & version	Final 1.0
Number of pages	18 (excluding front matter)
Contributing WP	2
WP/Task responsible	UNIGE
Other contributors	
Author(s)	Lonneke van der Plas, James Henderson, Paola Merlo
EC Project Officer	Philippe Gelin
Keywords	Semantic Decoder, Semantic Parsing

The partners in CLASSiC are:

Heriot-Watt University	HWU
University of Cambridge	UCAM
University of Geneva	GENE
Ecole Superieure d'Electricite	SUPELEC
France Telecom/ Orange Labs	FT
University of Edinburgh HCRC	EDIN

For copies of reports, updates on project activities and other CLASSiC-related information, contact:

The CLASSiC Project Co-ordinator:
Dr. Oliver Lemon
School of Mathematical and Computer Sciences (MACS)
Heriot-Watt University
Edinburgh
EH14 4AS
United Kingdom
O.Lemon@hw.ac.uk
Phone +44 (131) 451 3782 - Fax +44 (0)131 451 3327

Copies of reports and other material can also be accessed via the project's administration homepage, `http://www.classic-project.org`

Contents

Executive summary

This document describes the Prototype deliverable 2.3, due at month 18 of the CLASSIC project. The prototype is a syntactic-semantic parser for French, which has been built based on our existing syntactic-semantic parser for English, a syntactic Treebank for French, and a parallel, word-aligned English-French corpus. This document presents an overview of the steps we took to build the resources needed to train the syntactic-semantic parser for French, and its performance. Some of these steps produce prototypes of interest in their own right, so we also report results of evaluations on both the French syntactic parser trained on the existing Treebank and on the porting steps that map English annotation to French based on the parallel corpus. We also provide information on how to run the prototype.

To preview, we chose to use dependency parsing, rather than constituency parsing, because its word-based representation fits well with the word-based alignments we use for porting annotation across languages. We first developed a French syntactic parser by adapting our domain-general multilingual architecture for syntactic dependency parsing to French. Trained on the dependency version of the Paris French Treebank, we obtain state of the art results of 87.2% in labelled accuracy. With this parser, we parse the French side of the Europarl parallel English-French corpus (30-million words on each side). Also, using our previous model of syntactic-semantic dependency parsing for English, we parse the English side of the Europarl corpus. Because our previous models of dependency parsing were developed for the partially-artificial setting of the CoNLL shared tasks, both these parsing models required developing new components of the model to allow it to handle raw text as input. At the end of these two parsing steps, we have produced an English side with both syntactic and semantic annotation and a French side with syntactic annotation.

The two sides of this annotated parallel corpus are word-aligned automatically. We then port the English semantic annotation to the French side guided by the word-alignments. We now have both syntactic annotation and semantic annotation for the French sentences. We merge these two annoatations and the result is a French corpus with automatically generated syntactic-semantic annotation. After an optional filtering step, we train our syntactic-semantic parsing model on this artificial corpus. The current results for this parser are promising, but indicate that the ported SRL annotations are difficult for the current parsing model to learn, suggesting several directions for future improvement.

This report describes in more detail all these steps, the data on which the evaluation was performed, the larger scientific questions, and current work to improve mapping performance by a learning model. Some aspects of this work were published at ACL 2009 [11], at NAACL 2009 [1], at CoNLL 2009 [10], and at IJCAI 2009 [14].

1 Overview

Task 2.3 focuses on exploiting UNIGE's domain-general syntactic-semantic parsers and available syntactic-semantic annotations for English within a task to create training data for a French syntactic-semantic parser. This work has produced a working prototype. Some subtasks have performance surpassing the best existing comparable models.

This task provides enabling technology for French dialogue systems. It is designed to allow the Spoken Language Understanding module to handle greater syntactic and semantic complexity than that found in the TownInfo domain. This task will also provide annotated data for WP 6.

What in NLP is called Semantic Parsing is based on a task of Semantic Role Labelling (SRL). SRL consists of assigning abstract labels to participating individuals in a scenario-like frame expressed by a sentence. For example, given the sentence *She blames the government for her failure*, we want to label the noun phrases (*she, government*) as (JUDGE, EVALUEE) respectively, and disambiguate the prepositional phrase *for failure* as REASON. We also identify the verb *blame* as a verb of JUDGEMENT. Semantic role labelling has been shown to be useful in several user-interface tasks, such as question answering, and could provide an abstract representation to dialogue exchanges. Current state-of-the-art systems use supervised learning techniques to learn labels and, therefore, require large amounts of manually annotated data. Such data are available for English and a few other languages, but many languages remain that do not have large repositories of semantically annotated data. French is one such language.

Manual annotation is a labour-intensive and time-consuming enterprise. In line with our previous work [1] we investigate here a technique that consists of generating semantically annotated data for French automatically. We will port semantically annotated data from English to French using parallel word-aligned corpora.

The task of cross-lingual induction of such semantic role labels consists in projecting the semantic role labels based on an aligned parallel corpus. Similarity across languages has spurred recent interest in projecting automatically the semantic role labels from resource-rich languages to other resource-poorer languages, to automatically produce annotated data. This area of investigation has been explored for supervised bilingual projections from English to German and to a more limited extent to French by [2] in ground-breaking work, but much remains to be done. For example, the projections produced have not been used as training material for parsers. We apply here porting techniques to produce semantically annotated data for French with the aim of using the data to train a syntactic-semantic parser. So, for example, given the pair of sentences

She blames the government for her failure

and its French translation

Elle attribue son échec au gouvernement,

where the English side of the pair bears the Semantic Role Labels indicated below,

$$[_{JUDGE}\ She]\ blames\ [_{EVALUEE}\ the\ government]\ [_{REASON}\ for\ her\ failure],$$

we would like to automatically project the annotation to the French sentence obtaining

$$[_{JUDGE} Elle\ ] attribue\ [_{EVALUEE}\ au\ gouvernement]\ [_{REASON}\ son\ echec].$$

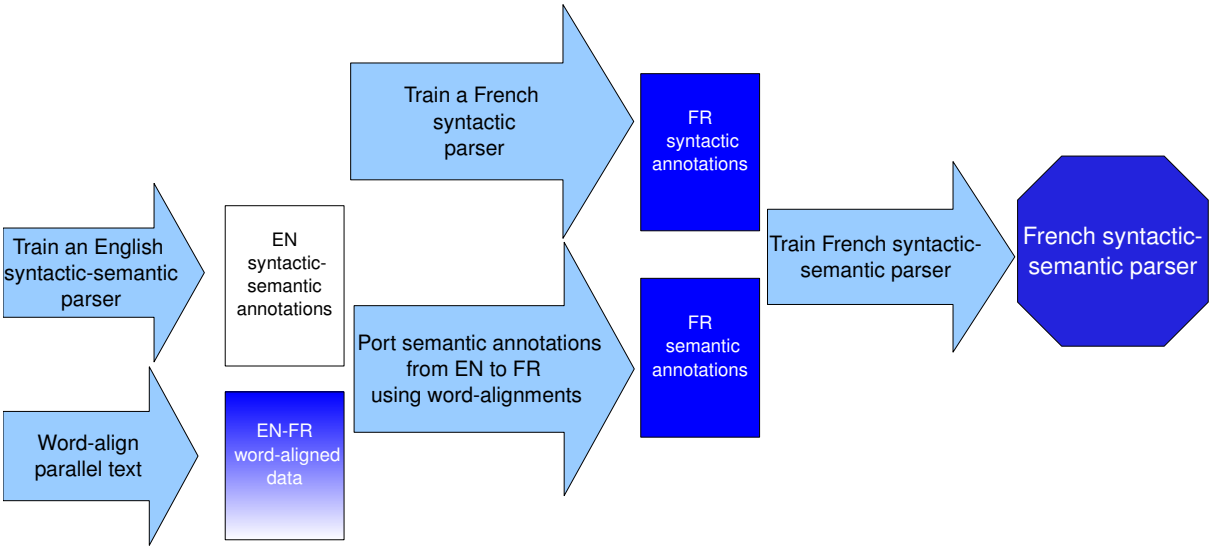

Figure 1: Overview of the porting process

In Figure 1, we can see an overview of the several steps we have taken to build a French syntactic-semantic parser. French components and resources are colour-coded in blue, English components and resources in white. We can see that we need both French syntactic annotation and French semantic annotation to train a syntactic-semantic parser for French. The syntactic annotation is acquired through the upper path. To compute French semantic annotations we have to follow the more complex lower part of the figure. For this we need a parallel corpus. Europarl [3] is a parallel corpus that includes English and French parallel texts. The English syntactic-semantic dependency parser described in Section 2 is used to annotate the English side of the parallel corpus. We word-aligned the English sentences to the French sentences automatically using GIZA++ [4]. We then ported the semantic annotation from the English side to the French side.

We break our presentation of this prototype and the related data into four subtasks: development, training and evaluation of an English statistical syntactic-semantic dependency parser (Section2); development, training and evaluation of a French statistical syntactic dependency parser (Section3); development and evaluation the French semantically annotated corpus (Section4); training and evaluation of the French statistical syntactic-semantic dependency parser (Section5).

2 Domain-general syntactic-semantic parsing for several languages

Recently, an increasing number of corpora have become available with shallow semantic annotations of written texts (for example, see the CoNLL 2008 and 2009 shared tasks). These semantic annotations are designed to be domain-general, so they should be a valuable resource for semantic interpretation tasks in a wide range of domains and for several languages. We have investigated exploiting such resources by first training a domain-general syntactic-semantic parser for English and then using the resulting parser to create a French corpus with semantic annotations. The same syntactic-semantic parsing model is then trained on this artificial French data, producing a syntactic-semantic parser for French.

For our syntactic-semantic parsing model we use the parser we developed last year for English in the CoNLL 2008 shared task. This syntactic-semantic dependency parser was extended this year to several languages with very good results [9, 10]. The good performance on multiple languages with a single model is particularly important in our task because we need a semantic parser that will also work well for French. We hypothesise that the use of a dependency representation for the semantic roles will be advantageous for porting across languages, because it fits better with the word-based alignments used in this porting process. Also, we use the PropBank database as a representation of semantic role labels. Based on results reported in [11], PropBank is more appropriate when syntactic and semantic representations are calculated jointly.

The crucial intuition behind the treatment of both syntax and semantic in a single model is that these two levels of information are related but not identical. We propose a solution that uses a generative history-based model to predict the most likely derivation of a synchronous dependency parser for both syntactic and semantic dependencies. Our probabilistic model is based on Incremental Sigmoid Belief Networks (ISBNs), a recently proposed latent variable model for syntactic structure prediction, which has shown very good behaviour for both constituency [12] and dependency parsing [13]. The ability of ISBNs to induce their features automatically enables us to extend this architecture to learning a synchronous parse of syntax and semantics without modification of the main architecture. By solving the problem with synchronous parsing, a probabilistic model is learnt which maximises the joint probability of the syntactic and semantic dependencies and thereby guarantees that the output structure is globally coherent, while at the same time building the two structures separately.

We devise separate derivations $D_d^1,...,D_d^{m_d}$ and $D_s^1,...,D_s^{m_s}$ for the syntactic and semantic dependency structures, respectively, and then divide each derivation into the chunks between shifting each word onto the stack, $c_d^t = D_d^{b_d^t},...,D_d^{e_d^t}$ and $c_s^t = D_s^{b_s^t},...,D_s^{e_s^t}$, where $D_d^{b_d^t-1} = D_s^{b_s^t-1} = shift_{t-1}$ and $D_d^{e_d^t+1} = D_s^{e_s^t+1} = shift_t$. The actions of the synchronous derivations consist of quadruples $C^t = (c_d^t, switch, c_s^t, shift_t)$, where *switch* means switching from syntactic to semantic mode. This gives us the following joint probability model, where n is the number of words in the input.

$$\begin{aligned} P(T_d,T_s) &= P(C^1,\ldots,C^n) \\ &= \textstyle\prod_t P(C^t|C^1,\ldots,C^{t-1}) \end{aligned} \tag{1}$$

The probability of each synchronous derivation chunk C^t is the product of four factors, related to the syntactic level, the semantic level and the two synchronising steps.

MODEL	CONLL MEASURES			CROSSING ARCS		
	Syntactic Labelled Acc.	Semantic F_1	Macro F_1	Semantics P	R	F_1
Johansson and Nugues 2008	89.3	81.6	85.5	67.0	44.5	53.5
UNIGE IJCAI 2009	87.5	76.1	81.8	62.1	29.4	39.9
UNIGE CoNLL 2008	87.6	73.1	80.5	72.6	1.7	3.3

Table 1: Scores on the test set.

$$\begin{aligned} P(C^t|C^1,\ldots,C^{t-1}) \quad = \\ P(c_d^t|C^1,\ldots,C^{t-1})\times \\ P(switch|c_d^t,C^1,\ldots,C^{t-1})\times \\ P(c_s^t|switch,c_d^t,C^1,\ldots,C^{t-1})\times \\ P(shift_t|c_d^t,c_s^t,C^1,\ldots,C^{t-1}) \end{aligned} \tag{2}$$

This model results in the performance shown in Table 1. We report both the official CoNLL 2008 shared task numbers [9] and the performances reported in [14], which are greatly improved through a better treatment of crossing arcs as described in [14]. We compare to the best performing model for English.

3 Development, training, and evaluation of a French statistical dependency parser

We already noted that syntactic annotation is available for French. The French Treebank [17] is a treebank of 21,564 sentences annotated with constituency annotation. A portion of the corpus (10,097 sentences) has been annotated with functional annotation as well. We use the automatic dependency conversion of the French Treebank into dependency format provided to us by M-H Candito and described in [18].

We used the parser described by [5], with the same division for training, testing, and development sets as in [6]: first 1235 sentences test, second 1235 sentences development, and the rest training.

We attain a 87.2% Labelled Attachment Score (LAS) on the test set. [18] trained a state-of-the-art constituency parser (the Berkeley parser [19]) and converted the output of that parser to dependency format as a post-processing step (using the same conversion rules that we used to create the corpus on which we trained the parser.) On the same test set, they have a labelled attachment score of 85% (unpublished results, Marie-Helne Candito, personal communication.) So our performance is 2% better than the previous work on this syntactic dependency parsing task.

The task defined by the CoNLL shared tasks is slightly artificial in that several pieces of information are provided as part of the input, including part-of-speech tags (automatically assigned). Thus both this syntactic parser and the syntactic-semantic parser discussed in the previous Section had to be augmented with a part-of-speech tagging component, as well as components for computing lemmas, and handling the lack of morphological feature information. POS tagging was added internally to the parsing models, in a manner similar to the constituent-based parser used for English in Task 2.2. This extension took a surprising amount of time. While often external POS tagging of the data yields slightly better performance, a full joint model of syntax (-semantics) and part-of-speech tagging results in more streamlined software, easier to use and to adapt, easier to apply to new data with different tagsets, and more in keeping with the CLASSiC architecture objective of components which are fully probabilistic.

The resulting trained syntactic dependency parser for French was then used to parse the French side of the Europarl data, discussed below. These parsed data are used in Section 5 below.

4 Building the French semantically-annotated corpus by porting annotation

Data-driven induction of semantic annotation based on parallel corpora is a well-defined and feasible task, and it has been argued to be particularly suitable to semantic role label annotation (as apposed to syntactic annotation) because cross-lingual parallelism improves as one moves to more abstract linguistic levels of representation. While [15, 16] find that direct syntactic dependency parallelism between English and Spanish fits 37% of dependency links, [2] reports an upper-bound mapping correspondence calculated on gold data of 88% F-measure for individual semantic roles, and 69% F-measure for whole scenario-like semantic frames.

4.1 The mapping hypothesis

Our mapping hypothesis is for the moment very simple, and no learning is used. We adopt a strong hypothesis of correspondence between languages [16].

> **Direct Semantic Correspondence Assumption (DSCA)**: Let a pair of sentences E and F that are (literal) translations of each other be given, with trees T_E and T_F. If vertices x_E and $y_E \in T_E$ are aligned with vertices x_F and $y_F \in T_F$ and if (syntactic-)semantic relationship $R(x_E, y_E)$ holds in T_E, then $R(x_F, y_F)$ holds in T_F.

We apply this hypothesis to semantic role dependency graphs, so the vertices are words and the relationship are labeled semantic roles. We also apply it to predicate sense labels, which refer to individual words. Currently we do not make any assumption about the correspondence between the syntactic annotations of the two languages.

As reported by [16], this is a strong hypothesis that is useful to trigger a process of role projection, but will not work correctly for several cases. We foresee a filtering step to filter out incorrect projections [16, 2].

4.2 Intersective word-alignment

For a pair of sentences, an alignment is a function mapping each word in the target language sentence to a word in the source language sentence. Alignments reflect co-occurrences in a corpus of translated sentence pairs, and are the basis of statistical machine translation. They can be found using efficient EM algorithms such as that used in GIZA++ [4]. The intersection of the word-alignments found in the source to target alignment and in the target to source alignment is called an intersective alignment. We have chosen to include only intersective alignments, because they have been reported to give the best results in machine translation and are, therefore, likely to support our cross-linguistic task better. They are also more likely to introduce recall errors rather than precision errors. We believe that when automatically creating annotated resources it is better to prefer high-precision annotation at the expense of coverage to avoid propagation of mistakes.

4.3 The porting algorithm

Following the Direct Semantic Correspondence Assumption, we port semantic annotation from English to French based on word alignments in the following way:

> For any pair of sentences E and F that are translations of each other, we port the semantic relationship $R(x_E, y_E)$ to $R(x_F, y_F)$ if and only if there exists a word-alignment between x_E and x_F and between y_E and y_F, and we port the semantic property $P(x_E)$ to $P(x_F)$ if and only if there exists a word-alignment between x_E and x_F.

The relationships which we port are semantic roles and the properties are predicate senses.

4.4 Evaluation framework for porting semantic information

We need to develop gold annotation of semantic labellings for French to evaluate the automatically ported semantic annotation. Following current practice in several conversions, for the moment we have obtained gold hand-annotated data annotated according to the FrameNet convention [7] that we have automatically converted to PropBank annotations [8] using the tools developed for the CoNLL '09 shared task by Sebastian Padó and Yi Zhang.[1] Although FrameNet is known to have a lower coverage than PropBank and the conversion is not flawless, we consider this to be an acceptable intermediate evaluation framework that will be validated in the second half of the year.

4.5 Experiments

We use the GIZA++ [4] implementation in the package for statistical machine translation Moses [30] to find intersective word alignments for the 1076 French sentences and their English counterparts. We parsed the 1076 English sentences using the parser described in Section 2. We ported the semantic annotation from the English sentences to the French sentences guided by the word alignments, as described in subsection 4.3.

The coverage of our porting process is affected by the intersective alignment method which we have chosen. It is a rather strict criterion, and, therefore, leaves many words non-aligned and their semantic annotation un-ported. Nonetheless, we were able to port most of the semantic annotation. We manage to port 78.0% of all predicates that are found in the English sentences (which are the output of the English parser). Furthermore, we managed to port 68.8% of all semantic roles.

Let us now take a look at results from evaluating on the gold sentences. 84.4% of the predicates in the gold sentences are found in the output of our system (i.e. unlabelled[2] predicate recall). Because the gold annotations are known to be incomplete due to the automatic conversion from FrameNet, it is not surprising that only 16.1% of predicates that are outputted by our system are found in the gold set. We, therefore, do not consider this predicate precision measure to be meaningful. For the same reason, we exclude from our semantic role performance measures the semantic roles for all output predicates for which there is no associated predicate in the gold data. The semantic role precision for the remaining

[1] Sebastian Pado kindly provided us with us with the FrameNet gold annotation of 1076 French sentences from Europarl. These are the gold data cited in [2].

[2] We are not able to determine whether the labels are correct, because the ported labels are for English and the gold labels are for French.

predicates (the percentage of ported semantic role relationships that are correct according to the gold) is 71.1%. The semantic role recall for the remaining predicates (the percentage of gold semantic role relationships that are found in the ported data) is lower, at 45.0%.

5 Training and evaluation of the French statistical syntactic-semantic dependency parser

In the previous section, we described how we ported semantic annotation from sentences in the source language (English) to word-aligned sentences in the target language (French) for a large parallel corpus. In Section 3, we described how a syntactic dependency parser for the target language was trained, and used to annotate the target side of the parallel corpus. We merged these two annotations to produce syntactic-semantic dependency corpus for the target language. In this section, we will report on using this corpus to train a syntactic-semantic parser for French, the end-product of Task 2.3.

5.1 Experiments

As described above, we word-aligned the parallel English-French Europarl corpus using GIZA++ [4] and selected intersective alignments. We applied minimal pre-processing to the corpus, following standard practice, to speed up the alignment process: we removed all sentences of more than 40 words. After removing empty lines and sentences longer than 40 words, we were left with 982,629 aligned sentence pairs. We parsed the English sentences with the syntactic-semantic parser trained on merged Penn Treebank, PropBank and ported the resulting semantic annotation to the French sentences. The syntactic annotation for the French sentences was computed by parsing the sentences with the French dependency parser described in Section 3. We merged the ported semantic annotation and this syntactic annotation to produce a syntactic-semantic annotation of the French side of the corpus. After removal of test and development sets of 1000 sentences each, we used these ported data to train syntactic-semantic parsers.

5.2 Initial results

Our initial experiments training on the ported data demonstrated how difficult this task is. Testing on the development set (which was also produced with the same porting procedure) after 4 iterations through the 980,602 training sentences, the syntactic labeled accuracy reached an impressive 88%, and the SRL precision (which includes both predicates and roles) reached 67%. But the SRL recall reached only 9%, meaning that the parser was outputting very few semantic role relationships. We hypothesise that this is due to noise in the ported training data, which introduces spurious predications for many words. This noise is impossible for the parser to predict, so it almost always chooses the simplest alternative of labeling a word as not a predicate, which in turn prevents that word from having any associated roles.

This analysis of our initial results suggests that better recall could be achieved by filtering out the sentences which are the source of this noise. We investigate this approach in the next subsection.

5.3 Improving the ported data using filters

Europarl is a very large corpus. Filtering noisy data is attractive because we can remove a large percentage of the training sentences and still have plenty of training data. In this section, we report our initial exploration of this approach, and the results of a relatively successful parser trained on the filtered data.

We have looked at three different ways of filtering the data. The first approach tries to directly address the problem identified above by filtering out predicates which only rarely appear with a given word. The second approach tries to identify errors by constraining the Part-of-Speech (PoS) of the candidate target predicate. For example, if the predicate in the source language is aligned to a predicate in the target language that is a determiner the predicate nor any of the accompanying roles gets ported. The third approach attempts to filter non-literal translations by imposing complete porting of all predicates and accompanying semantic roles in a sentence.

Filtering low frequency semantic annotation

Noise in the porting process and the very large data set means that a large number of words (by type) which should never be predicates end up being labeled as predicates a small number of times (by token) in the data set. Words which should be predicates also sometimes receive spurious predicate sense labels. In contrast, true predicate labels tend to occur frequently with a given word, because words do not have very many different senses. Motivated by these considerations, we removed a sentence from the data if it contained a predicate sense label with low relative frequency given its word.

We trained a syntactic-semantic parser on the remaining French Europarl data after removing all sentences which contain a *word-sense* pair where the relative frequency of *sense* given *word* is less than 0.2. This filter removed 78% of the sentences. When tested on the development set filtered in the same way (using frequencies from the training set), this parser reached around the same syntactic labeled accuracy (88%) and the same SRL precision (67%) as training on the full data set, but with an SRL recall of 42% rather than 9%. However, to get truely comparable evaluation results we need test this parser on the un-filtered development set. Here again the syntactic labeled accuracy is unchanged, but the SRL precision and recall go down to 49% and 28%, respectively. This result represents a better balance of precision to recall, and a better overall score of 63% (versus 60%), but it clearly leaves room for improvement.

We are investigating other filtering techniques (discussed below), but we have not yet trained parsers on those data. Thus, currently this is our best trained syntactic-semantic parser for French. This parser is the one we submit as our prototype. Work on improving these results will continue under Task 2.5.

Filtering semantic annotation based on PoS tags

The noise in the raw ported data discussed in Section 5.2 can also be used to motivate other filters which try to remove mistakes in the predicate labels of the ported data. In the English data, predicate labels are only assigned to verbs and nouns, because of the annotation guidelines. In general, not all Parts-of-Speech (PoS) can be predicates. For example, a determiner can never be a predicate. We investigated filtering predicates based on the PoS assigned to the word.

For all ported predicates in the development set we determined the PoS. In Table 2, we can see the number of times specific PoS labels are found attached to one of the ported predicates. The tag set used is that of Treebank+ [6]. The complete tagset is included at the end of this document.

Most of the candidate predicates are nouns (NC, NPP, ET). Many are verbs (V, VPP, VINF, VPR, VS,

1410	NC	5	VS
538	V	4	PRO
454	VINF	3	VIMP
211	VPP	3	ET
129	ADJ	1	PONCT
38	ADV	1	CS
33	P	1	CLS
30	VPR	1	CLR
17	NPP	1	CLO
13	DET		

Table 2: PoS of ported predicates on development set.

1200	NN
659	VB
453	NNS
405	VBN
315	VBP
272	VBG
167	VBZ
90	VBD

Table 3: PoS of predicates in English sentences of development set.

VIMP). However, many are also adjectives or adverbs (ADJ, ADV), prepositions (P), pronouns (PRO), and we even find some determiners (DET). Finally, we find clitics (CLS, CLR, CLO) a subordinating conjunction (CS) and a punctuation tag (PONCT).

If we run the same test on the English sentences we get the numbers described in Table 3. All predicates are either attached to nouns or verbs.

We, therefore, decided to construct the following filter: Do not port predicates and their roles to predicates in the target language if these candidate predicate carry a PoS other then *verb* or *noun*.

However, this filter seemed a bit strict in certain cases. The past participle in English is often tagged as an adjective in French. This adjective is perfectly capable of being a predicate. We, therefore, decided to construct a second filter. It ports predicates to predicates in the target language in the cases where the PoS of that candidate predicate is either a noun, a verb, or an adjective.

In Table 4 we see the result of filtering. Without any filtering 65.8% of semantic relations in the source text are ported. If we choose to port semantic annotation only for verb or noun predicates we lose some semantic annotation. 61.1% of semantic relations are ported. If we apply the filter that accepts adjectives we lose less data: 64% of semantic relations are ported.

Does this mean that many sentences will now have no semantic annotation at all? If we use no filter, 92% of the sentences have semantic annotation. If we choose to port semantic annotation only for verb or noun predicates 90% of the sentences carry some semantic annotation. If we apply the second filter that accepts adjectives we lose less data: 91% of the sentences carry semantic annotation.

The filters do not filter out many semantic relations and the number of sentences with semantic annotation

remains large. This might be positive, but the filter might also not be restrictive enough. Unfortunately, we are not able to measure the quality of the filter on the test set, because we have no PoS information for the test set, and assigning PoS tags is complicated by the presence of multi-word-units in these data.

A drawback of this approach to filtering is that the PoS tags are not gold information, the approach relies on the output of the PoS tagger applied and the tags might be wrong.

	No filter	verb + noun	verb + noun + adjective
% semantic rels	65.8	61.1	64.0
% sentences with SRL	91.8	89.9	91.1

Table 4: Results of filtering on PoS for the development set.

Filtering non-literal translations

Literalness of translations is often described as word-for-word translation. The less freedom given to the translator to render the text, the more literal the translation is likely to be. Non-literal translations are a common source of error in the projection of semantic roles from one text to a text in another language [2].

It is, therefore, not surprising that filtering mechanisms have been put in place to remove instances of non-literal translations from the parallel corpora being used. For example, [2] has applied translational consistency to filter out translational divergence by detecting rare patterns, variance in translation. [32] has proposed a measure of translational consistency based on information theory. To improve the quality of a machine translation system, [33] have used a bilingual dictionary to determine literalness of translations in a parallel corpus.

We impose a much weaker requirement than word-for-word translation by exploiting our semantic representation. Rather than requiring that all words in the source sentence are aligned to a word in the target sentence, we require that all words that are part of the semantic representation in the source text are aligned to a word in the target text. In other words, we remove target sentences from our corpus if the semantic annotation is not completely ported. In addition to this complete-porting filter, which requires all predicates and all roles from the source language to be mapped to the target language through word-alignments, we have also investigated a partial-porting filter. The partial-porting filter requires that all predicates are ported, but roles may still be unaligned and, therefore, not be ported.

In Table 5 we can see what happens to the data when the filters are applied. Without any filtering 91.8% of sentences carry semantic role annotation. If we choose to discard sentences for which the predicate and all of its roles are completely ported, we lose many sentences: only 22.9% of the sentences in the development set carry semantic annotation. If we apply the partial-porting filter we lose less data: 49.3% of the sentences carry semantic annotation. We see the same effect for the percentages of semantic relations that are ported. We go from 65.8%, when no filtering is applied to 14.6%, when the complete-porting filter is applied and 35.9% when the partial-porting filter is applied.

	No filter	Complete	Partial
% semantic rels	65.8	14.6	35.9
% sentences with SRL	91.8	22.9	49.3

Table 5: Results of filtering on completeness for the development set.

A large proportion of the data is discarded, but does the quality of the data improve? We compared the quality of semantic annotation resulting from the two filters on the test set. Results are given in Table 6.

	No filter	Complete	Partial
Precision	71.1	74.7	69.6
Recall	45.0	58.6	45.6

Table 6: Precision and recall when filtering on completeness for the test set.

The complete-porting filter clearly improves the data, especially in terms of recall. It seems that the partial-porting filter slightly deteriorates the quality of the data in terms of precision.

6 Conclusions

This task has achieved its main objectives. A prototype (D2.3) domain-general syntactic-semantic parser for French was completed on schedule. We consider this method to be a good baseline, and we see several directions for improvement, including excluding training sentences where the porting appears not to be successful, and making use of syntactic constraints to improve the word alignments used for porting. Further work is anticipated on this topic, to pursue these directions and to incorporate the new parsing models being developed in Task 2.5 for learning cross-lingual regularities. We will also exploit the trained models in Spoken Language Understanding as part of Task 2.4.

7 Running the Prototype Syntactic-Semantic Parser for French

The prototype consists of our synchronous model of syntactic-semantic parsing trained on the ported data filtered with the relative frequency sense filter, described in section 5.3. It has been tested under Linux. After unpacking `D2.3_prototype.tar.gz` by running

```
tar -zxf D2.3_prototype.tar.gz
```

enter the directory `D2.3_prototype/src/` and run

```
make
```

(This requires that you have the GNU Scientific Library (GSL) installed.) Then move back up to the directory `D2.3_prototype/`, where there should now be an executable `synsem_french_parser`. From here you can run the prototype interactively by running the script

```
./interactive.sh
```

Simply type sentences, one per line, with spaces separating clitics, d', l' and punctuation, as illustrated in `examples.snt`. The resulting parses will be displayed in the following format,

word-id word PoS head-id head-label predicate { *argument-id argument-role* }*

where "head" is the parent in the syntactic dependency graph, and the semantic role arguments for a given predicate are listed with the predicate. This script can also be run by piping sentences through it, as in

```
cat examples.snt | ./interactive.sh > examples.res
```

As illustrated inside the above scripts, the prototype can be invoked directly from the `model/` directory as

```
../synsem_french_parser -parsetext model1.par
```

In this mode, the output is produced in our ".ext" format, which can be converted to the official CoNLL 2009 format with

```
cd scripts; ./ext2conll2009 ../model/deps.nonproj ../input_file ../output_file
```

The latter script also converts from projectivised syntactic dependencies back to the original non-projective syntactic dependencies used in the CoNLL 2009 data. Parameters, such as the beam width used by the syntactic-semantic parser for decoding, can be adjusted by changing the number which follows the parameter in `model/model1.par`, or by adding"PARAM=value" to the end of the above command, as in:

```
./synsem_french_parser -parsetext model/model1.par BEAM=10
```

Parameters can also be used to run the parser on files, with "TEST_FILE=*input_file*" and "OUT_FILE=*output_file*".

8 Tagset Treebank+

In the Treebank+ [6] the following PoS tags are distinguished:

1. V (verb (indicatif))
2. VIMP (verb (impératif))
3. VINF (verb (infinitif))
4. VS (verb (subjonctif))
5. VPP (verb (participe passé))
6. VPR (verb (participe présent))
7. NPP (proper noun)
8. NC (common noun)
9. CS (subordinating conjunction)
10. CC (coordinating conjunction)
11. CLS (weak clitic pronoun (subject))
12. CLO (weak clitic pronoun (object))
13. CLR (weak clitic pronoun (reflexive))
14. P (preposition)
15. P+D(preposition)
16. P+PRO (preposition)
17. I (interjection)
18. PONCT (punctuation mark)
19. ET (foreign word)
20. ADJWH (WH adjective)
21. ADJ (adjective)
22. ADVWH (WH adverb)
23. ADV (adverb)
24. PROWH (WH strong pronoun)
25. PROREL (REL strong pronoun)

26. PRO (strong pronoun)
27. DETWH (WH determiner)
28. DET (determiner)

Bibliography

[1] L. van der Plas, J. Henderson, and P. Merlo. Domain adaptation with artificial data for semantic parsing of speech. In *In Proceedings of NAACL*, 2009.

[2] S. Padó. *Cross-lingual Annotation Projection Models for Role-Semantic Information*. PhD thesis, Saarland University, 2007.

[3] P. Koehn. Europarl: A multilingual corpus for evaluation of machine translation. 2003.

[4] F.J. Och. GIZA++: Training of statistical translation models. Available from `http://www.isi.edu/~och/GIZA++.html`, 2003.

[5] I. Titov and J. Henderson. A latent variable model for generative dependency parsing. In *Proceedings of the International Conference on Parsing Technologies (IWPT-07).*, 2007.

[6] B. Crabbé and M.-H. Candito. Expériences d'analyses syntaxique statistique du français. In *Actes TALN 2008*, 2008.

[7] C.F. Baker, C.J. Fillmore, and J.B. Lowe. The Berkeley FrameNet project. In *In Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics (ACL-COLING'98)*, 1998.

[8] M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:31:71–105, 2005.

[9] J. Henderson, P. Merlo, G. Musillo, and I. Titov. A latent variable model of synchronous parsing for syntactic and semantic dependencies. In *Proceedings of CoNLL 2008*, pages 178–182, Manchester, UK, 2008.

[10] A. Gesmundo, J. Henderson, P. Merlo, and I. Titov. A latent variable model of synchronous syntactic-semantic parsing for multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 37–42, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[11] P. Merlo and L. van der Plas. Abstraction and generalisation in semantic role labels: Propbank, Verbnet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[12] I. Titov and J. Henderson. Constituent parsing with Incremental Sigmoid Belief Networks. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 632–639, Prague, Czech Republic, 2007.

[13] I. Titov and J. Henderson. A latent variable model for generative dependency parsing. In *Proceedings of the International Conference on Parsing Technologies (IWPT'07)*, Prague, Czech Republic, 2007.

[14] I. Titov, J. Henderson, P. Merlo, and G. Musillo. Online Graph Planarisation for Synchronous Parsing of Semantic and Syntactic Dependencies. In *Proc. Int. Joint Conferences on Artificial Intelligence (IJCAI-09)*, pages 1562–1567, Pasadena, CA, USA, 2009.

[15] R. Hwa, P. Resnik, A. Weinberg, and O. Kolak. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 392–399, 2002.

[16] R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. Bootstrapping parsers via syntactic projection across parallel text. *Natural Language Engineering*, 11(3):311–325, 2005.

[17] A. Abeillé, L. Clément, and F. Toussenel. Building a treebank for french. In *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers, 2003.

[18] M.-H. Candito, B. Crabbé, P. Denis, and F. Guérin. Analyse syntaxique du français : des constituants aux dépendances. In *Proceedings of TALN 2009*, 2009.

[19] S. Petrov, L. Barrett, R. Thibaux, and D. Klein. Learning accurate, compact, and interpretable tree annotation. In *Proc. of the Annual Meeting of the ACL and the International Conference on Computational Linguistics*, Sydney, Australia, 2006.

[20] A. Arun and F. Keller. Lexicalisation in probabilistic parsing: the case of french. In *Proceeding of the 43d annual meeting of the ACL*, 2005.

[21] P. Fung and B. Chen. BiFrameNet:Bilingual frame semantic resource construction by cross-lingual induction. In *Proceedings of COLING*, 2004.

[22] P. Fung, Z. Wu, Y. Yang, and D. Wu. Learning bilingual semantic frames: Shallow semantic parsing vs. semantic role projection. In *11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, 2007.

[23] R. Hwa, P.Resnik, A. Weinberg, and O. Kolak. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the ACL*, 2002.

[24] R. Hwa, Ph Resnik, A.Weinberg, C. Cabezas, and O. Kolak. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 1:1–15, 2004.

[25] R. Johansson and P. Nugues. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of the annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.

[26] S. Padó and M. Lapata. Cross-linguistic projection of role-semantic information. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[27] S. Padó and G. Pitel. Annotation précise du fran cais en sémantique de rôles par projection cross-linguistique. In *Proceedings of TALN*, 2007.

[28] M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Comp. Ling.*, 31:71–105, 2005.

[29] S. Petrov, L. Barrett, R. Thibaux, and D. Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the Annual Meeting of the ACL*, 2006.

[30] A. Birch and C. Callison-Burch and M. Federico and N. Bertoldi and B. Cowan and W. Shen and C. Moran and R. Zens and C. Dyer and O. Bojar and A. Constantin and E. Herbst and P. Koehn and H. Hoang. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007.

[31] D. Yarowsky, G. Ngai, and R. Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the International Conference on Human Language Technology (HLT)*, 2001.

[32] I.D. Melamed. Measuring semantic entropy. In *In Proceedings of the ANLP SIGLEX Workshop on tagging text with lexical semantics*, 1996.

[33] K. Imamura, E. Sumita, and Y. Matsumoto Automatic construction of machine translation knowledge using translation literalness. In *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 2003.