Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification

Chloé Braud
ALPAGE, Univ Paris Diderot
& INRIA Paris-Rocquencourt
75013 Paris - France
chloe.braud@inria.fr

Pascal Denis
MAGNET, INRIA Lille Nord-Europe
59650 Villeneuve d'Ascq - France
pascal.denis@inria.fr

Abstract

This paper presents the first experiments on identifying implicit discourse relations (i.e., relations lacking an overt discourse connective) in French. Given the little amount of annotated data for this task, our system resorts to additional data automatically labeled using unambiguous connectives, a method introduced by (Marcu and Echihabi, 2002). We first show that a system trained solely on these artificial data does not generalize well to natural implicit examples, thus echoing the conclusion made by (Sporleder and Lascarides, 2008) for English. We then explain these initial results by analyzing the different types of distribution difference between natural and artificial implicit data. This finally leads us to propose a number of very simple methods, all inspired from work on domain adaptation, for combining the two types of data. Through various experiments on the French ANNODIS corpus, we show that our best system achieves an accuracy of 41.7%, corresponding to a 4.4% significant gain over a system solely trained on manually labeled data.

1 Introduction

An important bottleneck for automatic discourse understanding is the proper identification of implicit relations between discourse units. What makes these relations difficult is that they lack strong surface cues like a discourse marker. This point is illustrated in the French examples (1) and (2).[1] In (1), the connective *mais* (*but*) triggers a relation of *contrast*, whereas in (2), there is no explicit connective to signal the *explanation* relation, and the relation has to be inferred through other ways (in this case, a causal relation between having injured players and loosing).

(1) La hulotte est un rapace nocturne, **mais** elle peut vivre le jour.
*The tawny owl is a nocturnal bird of prey, **but** it can live in the daytime.*

(2) L'équipe a perdu lamentablement hier. Elle avait trop de blessés.
The team lost miserably yesterday. It had too many injured players.

Implicit relations are very widespread in naturally-occurring data. Thus, they make up between 39.5% and 54% of the annotated examples in the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), depending on the relation types used.[2] A quick look at other discourse corpora suggests that the problem is as pervasive (if not more) in other languages. The French ANNODIS corpus does not annotate the distinction between explicit and implicit relations, but a projection of a French connective lexicon on the data gives a proportion of 47.4 to 71% of implicit relations, depending on the set of relations.[3] For the German discourse corpus of (Gastel et al., 2011), (Versley, 2013) report 65% of implicit relations.

In this paper, we tackle the problem of automatically identifying implicit discourse relations in French. To date, the large majority of studies on this task have focused on English, and to a lesser extent on German. Performance remain relatively low compared to explicit relations, due to the lack of strong

[1] All our examples are taken from the ANNODIS corpus: http://redac.univ-tlse2.fr/corpus/annodis/.

[2] The former count does not include *AltLex*, *EntRel* and *NoRel* as implicit examples, whereas the latter does.

[3] The first count does not include *attribution*, *e-elaboration* and *frame* examples.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1694–1705, Dublin, Ireland, August 23-29 2014.

predictors. Because it relies on more complex, interacting factors, the identification of implicit relations requires a lot of data. But the available annotated for French is scarce: while the PDTB contains about $40,000$ examples, the French ANNODIS only has about $3,000$ examples. An additional challenge for building such a system for French compared to English is the lack of external lexical ressources (e.g., semantic verb classification, polarity database).

A natural approach to deal with the lack of annotated implicit data is to resort to additional data automatically obtained from explicit examples in which the connective is removed (Marcu and Echihabi, 2002). Provided that one could reliably identify discourse connectives, this approach makes it possible to create large amounts of additional implicit data from raw texts. Unfortunately, (Sporleder and Lascarides, 2008) show that a system trained on this type of artificially generated data does not generalize well, leading to important performance degradation compared to a system solely trained on natural data.

The central question we address in this paper is how to better leverage the large amount of automatically generated data. We first show that the bad generalization performance of the system trained on artificial data lies in important distribution differences between the two datasets. This analysis in turn leads us to investigate various simple schemes for combining natural and artificial data methods inspired from the field of domain adaptation. Our best combined system yields a significant improvement of 4.4% over a system solely trained on the available manually annotated data.

The rest of this paper is organized as follows. Section 2 summarizes previous works on implicit relation identification. In section 3, we describe the problems introduced by the use of artificial data and the methods we develop to deal with them. In section 4, we give a description of the data used, and in section 5, we detail our feature set. Our experiments are then summarized in section 6.

2 Related Work

To date, there have been only a few attempts at building full document-based discourse parsers. On the RST-DT (Carlson et al., 2001), the best performing system is (Joty et al., 2013), who report an F_1 score of 55.71 for labeled stuctures (with 23 relations). On the same corpus, (Sagae, 2009) and (Hernault et al., 2010) report F_1 scores of 44.5 and 47.3, respectively. On the PDTB, the parser of Lin et al. (2010) obtains an F_1 score of 33 (16 explicit relations, 11 implicit relations). On the ANNODIS corpus, Muller et al. (2012) reports F_1 scores of 36.1 (17 relations) and 46.8 (4 relations).

These still modest performance are due to wrong attachment decisions, as well as to errors in relation labeling. Most of these latter errors are mostly imputable to wrong classifications of implicit relations. Thus, the current best accuracy performance on explicit PDTB relations are 94.15% on 4 relations (Pitler and Nenkova, 2009), and 86.77% on 16 relations (Lin et al., 2010). By contrast, the best identification system for implicit PDTB relations obtains an accuracy of 65.4% on 4 relations in (Pitler et al., 2009), and down to 40.2% for 11 of the level 2 relations of PDTB (Lin et al., 2009). For German, Versley (2013)'s study on implicit relations reports 42.5 in F_1 for 5 relations and 18.7 for 21 relations. For French, Muller et al. (2012) report an accuracy score of 63.6% for their relation labeling system (over 17 relations), but they do not provide separate scores for explicit vs. implicit relations.

This performance drop reflects the difficulty of identifying a rhetorical relation in the absence of an explicit discourse marker. As shown by (Park and Cardie, 2012), the identification of implicit relations relies on more diverse and noisy predictors from syntax (in the form of prediction rules) and (lexical) semantics (e.g., polarity, semantic classes and fine-grained semantic tags for verbs). Unfortunately, most of the semantic resources used to derive features for English (polarity database, Inquirer tags) are not available for French. Zhou et al. (2010) try to predict the implicit connectives annotated in the PDTB as a way of predicting the relation, a method only possible with this corpus. They obtain results lower than those reported by (Park and Cardie, 2012). In another context, Sporleder (2008) shows that using WordNet is less effective than lemmatisation for capturing semantic generalization, and (Wang et al., 2010) use tree kernels in order to better capture important syntactic information. In another context, Sporleder (2008) shows that using WordNet is less effective than lemmatisation for capturing semantic generalization, and (Wang et al., 2010) use tree kernels in order to better capture important syntactic information.

Another set of studies we directly build upon explore the idea that many connectives unambiguously trigger a unique relation, thus allowing to construct massive amount of (artificially) labelled implicit examples from raw data. Marcu and Echihabi (2002) were the first to use this method: they were mainly interested in showing that a removed connective could be recovered from its linguistic context. In turn, they only tested their approach on examples that were also generated automatically, and not on manually annotated implicit examples. In this setting, they report an accuracy of 49.7 (6 classes), significantly above luck. Reusing the same approach, Sporleder and Lascarides (2008) then showed that a system trained on a large amount of artificial examples (72000 examples) performs much worse than the same system trained on a much smaller amount of natural examples ($1,051$ examples) implicit examples, with accuracies of 25.8 and 40.3, respectively.

Marcu and Echihabi's (2002) original approach was based on the idea of finding pairs of semantically related words that together trigger a relation (such as "nocturne/jour" ("nocturnal/daytime") in example 1 of *contrast*). Interestingly, Pitler et al. (2009) showed that word pairs extracted from artificial data are not helpful for implicit relation identification and, moreover, that the most informative word pairs are not semantically related. Blair-Goldensohn et al. (2007) showed that, for *cause* and *contrast* at least, results can be enhanced by improving the quality of the artificial data. Finally, Wang et al. (2012) propose a first approach that exploits both natural and artificial data. Specifically, they select the most informative training points among natural and artificial examples, both coming from the PDTB or the RST DT. They define deterministic rules for identifying so-called "typical" examples of a relation, the "seed" sets that are then expanded using a simple clustering algorithm. They report performance results well over those of (Pitler et al., 2009), but using a different evaluation protocole. [4] Also, their method is not easy to repoduce, especially for French, where we can not define the same deterministic rules as some of these depend on polarity information, for which we do not have external resources. Furthermore, their approach only extracts 1 to 5% of the data as seed examples, which would represent too few examples on our corpus. Finally, we are interested in finer-grained relations, thus more difficult to discriminate using these kind of rules.

3 Proposed Approach

Our approach builds upon and extends the method of (Marcu and Echihabi, 2002) and (Sporleder and Lascarides, 2008) by investigating different strategies for combining natural and artificial examples of implicit discourse relations. These different combination schemes are inspired from domain adaptation and are motivated by the fact that artificial and natural examples follow different probability distributions.

3.1 Distribution Differences

Most machine learning algorithms are based on the assumption that data from training and test samples are independently and identically distributed (i.e., the i.i.d. sampling assumption). Yet, it seems that the use of artificial data clearly undermines this assumption. There is indeed no guarantee that our artificial examples should follow a distribution similar to that of the manual examples. This leads to the problem of learning from non-iid data, a problem that has attracted growing attention these last years in machine learning and NLP (Sogaard, 2013), (Hand, 2006).

In this particular context, we have two sets of data with the same output space (i.e., the discourse relations), and the same kind of inputs space (i.e., spans of text). But our data samples can differ in a number of ways. Following the terminology in (Moreno-Torres et al., 2012), we may encounter all the different kinds of *shift* that can appear in a classification problem.

Prior Probability Shift This shift describes changes in the marginal distribution of the *output* (i.e., the relations). The artificial data do not have the same class distribution as the natural ones (see section 4). Neither do they have the same distribution as the natural explicit, because of the automatic extraction. This problem can be easily handled by resampling artificial data (see section 4).

[4] Wang et al. (2012) only use the first annotated relation and ignore the *Entity* relation, whereas Pitler et al. (2009) keep all the annotations and map *Entity* examples to the *Expansion* class.

Covariate Shift This shift describes changes in the marginal distribution of the *input* (i.e., the pairs of spans of text). Artifical examples are originally explicit examples minus their connective, so it is reasonnable to think that these examples will have a different distribution from the natural implicit examples. Moreover, it is possible that, by removing the connective, we have made these examples semantically unfelicitous or even ungrammatical. Segmentation is another issue, since it is automatic and based on heuristics (see section 4). For example, artificial examples can not be multi-sentential whereas it can be the case for natural ones.

Concept Shift This shift describes changes in the *joint* distribution of inputs and outputs. Consider for instance the occurrences of relations within inter- and intra-sentential contexts. The proportion of inter-sentential examples in natural and artificial datasets is the same for *contrast* (57.1%), it is similar for *result* (resp. 45.7% and 39.8%), but very different for *continuation* (resp. 70% and 96.5%) and for *explanation* (resp. 21.4% and 53.0%). Moreover, the extraction method is prone to errors, and it may be the case that we wrongly identify a word form as a discourse connective. Thus, we may produce examples annotated with a wrong relation or that do not involve any discourse relation at all. Finally, deleting a connective can make the discourse ackward or even incoherent (Asher and Lascarides, 2003). We can actually witness this with example (1). As shown by (Sporleder and Lascarides, 2008), deleting the connective can also change the inferred relation. They found examples of *explanation* in which an implicit relation becomes the only one inferable after removing the explicit marker.The deletion can also change the inferred relation (Sporleder and Lascarides, 2008). We found an even worse effect in our French corpus. In example (3), the connective *puisqu(e)* (*because*) triggers an *explanation*, thus the events are ordered following the causal law. The cause, "migrer" ("migrate"), comes before the effect, "deviennent" ("becomes"). But when we delete the connective, the order of the events seems to be reversed. Keeping the first clause as the first argument, we then obtain a *result* relation in this sentence.

(3) Les Amorrites deviennent à la période suivante de sérieux adversaires des souverains d'Ur, **puisqu**'ils commencent alors à migrer en grand nombre vers la Mésopotamie.
In the next period, Amorrites become severe opponents of the sovereigns of Ur, ***because*** *they then begin to migrate in large numbers to Mesopotamia.*

3.2 Methods Inspired by Domain Adaptation

A way to deal with all the distribution differences observed is to reframe our problem within the framework of domain adaptation. Informally, the task of domain adaptation is to port some system from one domain, the *source*, to another, the *target*. Informally, we have a distribution D_s for the source data and a distribution D_t for the target data. The goal of the classifier is to build a good approximation of D_t. If one uses data following the distribution D_s in order to build this approximation, then the performance will depend of the similarity between D_s and D_t. If these distributions are too dissimilar, the approximation will be bad and so will be the performance. It is the case in particular when the domains (e.g., text genres) are different. The goal of domain adaptation is precisely to deal with data from different distributions (Jiang, 2008), (Mansour et al., 2009). We are not exactly in the same setting, but we can regard the artificial data as the *source*, and the natural data, on which we evaluate, as the *target*.

As a first step, we decided to investigate the simplest domain adaptation methods there is, such as those described in (Daumé III, 2007). These methods either combine directly the data or the models built on each set of data. Performance of all these systems will be compared to the base systems trained on only one set of data, in section 6.

Data combination The first possibility is to combine the data. The first model is trained on all natural *and* artificial data together (UNION). This method does not allow us to control the importance of the two sets of data nor to evaluate their influence on the system. We thus refine it in two ways. First, we only add to the manual data randomly selected samples from the artificial data (ARTSUB). Alternatively, we keep all the artificial examples but reweight (or, equivalently, duplicate) the manual examples (NATW). Both these schemes allow us to avoid a massive imbalance between the two kinds of data.

Model combination The second strategy consists in combining the models. A first set of methods involve adding new features. That is, we train a model on the artificial data, then run it on the natural examples. We use these predictions as new attributes for the natural model (ADDPRED). The parameter associated to the attribute therefore measures the importance to be given to the predictions made by the model trained on artificial data. We propose a variation of this method by adding the probabilities of each prediction as supplementary attributes (ADDPROB). The intuition is that even if the classifier is wrong, it could still be consistent in its errors. Yet another model combination consists in using the parameters of the artificial model as initial values for the manual model parameters (ARTINIT). This method allows to give an initial information to the natural model rather than a random intialization. Finally, we also build a model by linearly interpolating the two basic models (LININT).

In addition to these combination schemes, we also add a method to automatically select examples among the artificial set based on the confidence of the artificial model. Its aim is to filter out noisy examples, our hypothesis being that the more confident the model, the less noisy the example.

4 Data

In this work, we choose to focus on 4 relations, *contrast*, *result*, *continuation* and *explanation*, each of which can be either explicit or implicit. These are the same as the relations used in (Sporleder and Lascarides, 2008), allowing for easy comparison across languages, with the exception of the relation *summary* which does not appear in the ANNODIS corpus. Although it is difficult to map these relations onto the relation set of the PDTB, we can say that our relations are closer to level 2 and level 3 (i.e., fine-grained) PDTB relations than level 1 (i.e., coarse-grained) ones.

4.1 Manually Annotated Data: ANNODIS

Our natural implicit examples are taken from the ANNODIS corpus, which is to date the only available French corpus annotated at the discourse level. Its annotations are based on the SDRT framework (Asher and Lascarides, 2003). It consists of 86 newspaper and Wikipedia articles. 3,339 examples have been annotated using 17 relations. In way of comparison, note that the PDTB has roughly 12 times more annotated relations than ANNODIS. Documents are segmented in Elementary Discourse Units (EDUs) which can be clauses, prepositionnal phrases and some adverbials and parentheticals if the span of text describes an event. The relations link EDUs and complex segments, adjacent or not. The connectives are not annotated, which means that the examples of implicit relations had to be extracted automatically.

The corpus has been pre-processed using the MELt tagger (Denis and Sagot, 2009) for POS-tagging, lemmatization and morphological markings. Then, the documents have been parsed using the the MST-Parser (McDonald and Pereira, 2006) trained for French by (Candito et al., 2010). In order to identify implicit examples, we used the French lexicon of connectives (LexConn) developed by Roze et al. (2012). We simply matched all possible connective forms associated with the annotated relations (discarding *à*, which is too ambiguous). We did not add constraints on the connective position, as we wanted to be sure to exclude all explicit examples, this method led us to miss a few implicit examples. Out of 1,108 examples annotated with one of the 4 relations considered, 494 were found to be implicit (see table 2).

4.2 Automatically Annotated Data

The artificial data are automatically extracted from raw data using heuristic rules. We use LexConn to mine explicit instances in the corpus Est Républicain composed of newspaper articles (9M sentences), with the same pre-processings as ANNODIS. LexConn contains 329 connectives, among them, 131 are unambiguous for our 4 relations. We grouped pragmatic relations (i.e., the relation is between speech acts) and non pragmatic relations (i.e., the relation is between facts) relations, assuming they involve the same kind of predictors, and the 3 contrastive relations, as only one type of *contrast* is annotated in ANNODIS. We did not take into account 3 connectives corresponding to unknown part-of-speech. Our first evaluation led us to delete 6 connectives, very ambiguous between discourse and non discourse readings, such as "maintenant" ("now"). We eventually settled on 122 connectives, among which 100 were seen in the corpus in a configuration matching one of our pre-defined patterns. As a comparison, (Sporleder

and Lascarides, 2008) only had 50 such connectives. We finally use 122 connectives, among which 100 were seen in a correct configuration in the corpus. As a comparison, 50 were used in (Sporleder and Lascarides, 2008).

Position	Part-of-speech	Patterns	Examples
Inter-sentential	All POS	A1. **C**(,) A2.	A1. **Malheureusement**(,) A2 A1. **Surtout**, A2.
	Adv.	A1. beg-A2(,) **C**(,) end-A2. A1. A2, **C**.	A1. beg-A2, **de plus**, end-A2. A1. beg-A2(,) **en outre**(,) end-A2. A1. A2, **remarque**.
Intra-sentential	All POS	A1, **C**(,) A2.	A1, **de plus**(,) A2. A1(,) **donc**(,) A2.
	SC and Prep.	**C** A1, A2.	**Preuve que** A1, A2. **Puisque** A1, A2.
	Adv.	A1, beg-A2(,) **C** (,) end-A2. A1, A2, **C**.	A1, beg-A2, **de plus**, end-A2. A1, beg-A2(,) **en outre**(,) A2. A1, A2, **réflexion faite**.

Table 1: Defined patterns with some examples. "A1" stands for the first argument, "A2" for the second and "**C**" stands for the connective ; "beg" and "end" stand resp. for the beginning and the end of an argument ; "(x)" indicates that "x" is not necessary, depending on the connective form. Some patterns are only possible for some sets of connectives based on their part-of-speech (Subordinating Conjunction (SC), Preposition (Prep.), Averbials (Adv.)).

The heuristic used to extract the examples has two main steps. First, we search forms used in discourse readings using patterns (see table 1) that were manually defined for each connective based on its position, its part-of-speechand the punctuation around it. Second, we identify the connectives arguments using the same information. We make the same simplifying assumptions as in the previous studies: an argument covers at most one sentence, and we have at most 2 EDUs within a sentence. As additional constraint, we also require the presence of a verb in each relation argument. When two connectives occur in the same segment, it is possible that one modifies the other. In turn, a naive extraction could produce two examples with different relations but the same arguments. To avoid the creation of spurious examples, we extract two examples in these cases only if one is inter- and the other intra-sentential according to our extraction patterns.

	Natural dataset		Artificial dataset		
Relation	Explicit	Implicit	Available	Training	Test
contrast	100	42	252 793	23 409	2 926
result	52	110	50 297	23 409	2 926
continuation	404	272	29 261	23 409	2 926
explanation	58	70	59 909	23 409	2 926
All	614	494	392 260	93 636	11 704

Table 2: Number of examples in our corpora, for the natural dataset, only the implicit examples are used.

This simple method allows to quickly generate a large amount of data. In total, we extracted $392,260$ examples (see table 2). This initial dataset was rebalanced in a way to keep the maximum number of available examples (thus dealing with the prior probability shift). We used 80% of the data as training set, and 10% the development and test set. Note that there are some important differences in the label distributions between natural and artificial data. For instance, the most represented relation in the natural data (*continuation*) is the least represented in the artificial data. This is because the connectives that trigger this relation are highly ambiguous between discourse and non-discourse readings. Finally, this method generates some noise: out of 250 random examples, we found 37 errors in span boudaries and

18 cases in which the connective form does not have a discourse reading.

5 Features

We adapted various features used in previous studies. The lack of ressources for French prevented us from using them all, especially the semantic ones. These features correspond to surface information and others more linguistic. As a comparison, (Marcu and Echihabi, 2002) only used pairs of words.

Sporleder and Lascarides (2008) used various linguistic features but no syntaxic ones. (Wang et al., 2012) used semantic, syntaxic and lexical information. We used lexico-syntaxic information. Finally, note that our goal is to evaluate the efficiency of data combinations. Thus we did not try to optimize this feature set, as it would have introduced another parameter in our model.

Indication of syntactic complexity: we compute the number of nominal, verbal, prepositional, adjectival and adverbial phrases.

Information concerning the heads of the arguments: we keep the lemma of negative element linked to the head, we also get some temporal/aspectual information (number of auxiliaries dependent of the head, tense, person, number of the auxiliaries), information about the heads dependents (if an object, a by-object or a modifier is present ; if a preposition dependent of the head, subject or object is present ; part-of-speech of the modifiers and prepositional dependents of the head, subject and object) and some morphological information (tense and person of the head if verbal, gender if non verbal, number of the head, precise part-of-speech, "VPP", and simplified,"V"). We also add features pairing the tenses for verbal heads and the heads numbers.

Position: we add a feature indicating if the example is inter or intra-sentential.

Indication of thematic continuity: we compute general lemma overlap and lemma overlap for open class words.

6 Experiments

Our main objective is to assess whether one can use the artificial data to improve the performance of a system solely based on data manually annotated only available in small amount. We therefore test the methods described in section 3.

We experimented with a maximum entropy classifier from the MegaM[5] package, in multiclass classification, with a maximum of 100 iterations. We did not try to optimize the regularization parameter which is then equal to 1.

We rebalance the corpus of manually annotated data to a maximum of 70 examples per relation.[6] We have too few annotated examples to be able to construct a separate test set sufficiently large to make statistical significance test. Thus, we decided to make a stratified nested cross-validation. It has been shown that this method provides an estimate of the error that is very close to that one could obtain on an independent evaluation set ((Varma and Simon, 2006), (Scheffer, 1999)), as it prevents us from optimizing our hyper-parameters and performing evaluation on the same data. Specifically, there are two cross-validation loops: the inner loop is used for tuning the hyper-parameters (as described in section 6.2) and the outer loop estimates the generalization error. The data are first split into N folds. We take the fold k (with $1 \leq k \leq N$) as the current evaluation set. The $N-1$ other folds are used as training data and split into M folds used for model fitting. The best model is then evaluated on the fold k. Finally, we report performance on the N folds. We used two 5-fold cross-validation in order to select and evaluate the best models for the systems described in section 3.2. We have no guarantee to select the best models at each test step, but this procedure allows to evaluate the stability of the system with respect to the hyper-parameters (i.e. the chosen values should not be too scattered), the overfitting (i.e. inner and

[5] http://www.umiacs.umd.edu/~hal/megam/version0_3/

[6] Our focus is on the methodology of data combination, so we left for future work the issue of dealing with the highly imbalanced relation distribution of the natural data. Incidentally, note that this setting prevents us from getting a system solely performing well on highly frequent relations.

outer estimations should be close) and the stability of the models (i.e. variance in the predictive capacity, between the results on the outer folds).

As in the previous studies, we report performance using micro-averaged accuracy and F_1 score per relation. In order to evaluate the statistical significance of our results, we use the Student's t-test (with p-value < 0.05) which has been proved to work with very small sample (see (de Winter, 2013)) if the effect size (computed using the Cohen coefficient) and the correlation between the sample are large enough, while, as noted in (de Winter, 2013), the Wilcoxon signed rank test (that we initially tried) could lead to overestimated p-value with such small sample. The results of the most relevant systems are presented in table 3.

	Without selection				With selection	
	NATONLY	ARTONLY	ADDPRED	ARTINIT	ADDPRED+SELEC	NATW+selec
Accuracy	37.3	23.0	39.3	40.1	**41.7***	41.3
contrast	15.0	23.2	16.0	16.9	20.8	19.2
result	47.6	15.7	50.6	45.9	51.0	48.3
continuation	28.1	32.1	31.9	34.0	31.2	32.4
explanation	47.9	22.4	46.7	52.2	53.9	53.4

Table 3: Most relevant systems, with or without selection of examples, overall accuracy and F_1 score per relation, * corresponds to a significant improvement over NATONLY.

6.1 Basic Models

In the first set of experiments, we trained two classifiers. The first one is trained on the natural implicit data (NATONLY, 252 examples), and the second one on the artificial implicit data (ARTONLY, 93,636 examples). We test both models on natural implicit data.

The overall accuracy of the NATONLY model is 37.3 with F_1 score ranging from 15.0 for *contrast* to 47.9 for *explanation*. The performance on *contrast* is fairly low, probably because this relation is the least frequent in our training set. Note that the overall accuracy obtained is quite close to the 40.3 obtained for English by (Sporleder and Lascarides, 2008).

The overall accuracy of the ARTONLY model is 47.8 when evaluated on the same type of data, that is, artificial ones (11,704 test examples), but only 23.0 when evaluated on natural data. This significant drop in performance has been observed in the previous studies on English. It can be attributed to the distribution differences described in section 3. We can observe that the use of the artificial data lowers the F_1 score for *result* and *explanation* while, for *contrast*, F_1 score is raised by about 10 points.

6.2 Models with Combinations

In this section, we present the results for the systems using both natural and artificial data. We either directly combine the data or use the data to build separate models that are then combined. Some of these models use hyper-parameters. When weighting the natural examples, we test weights $c \in [0.5, 1, 5]$ and $c \in [10; 2000]$ with an increment of 10 until 100, of 50 until 1000 and of 500 until 2000. When adding random subsets of artificial data, we add each time k times the number of natural examples artificial examples with $k \in [0.1; 600]$ with an increment of 0.1 until 1, of 10 until 100 and of 50 until 600. Finally, when taking a linear interpolation of the models, we build a new model by weighting the artificial model by $\alpha \in [0.1; 0.9]$ with increments of 0.1.

In general, we observe that most of the systems lead to similar or higher accuracy than NATONLY, but none of the improvements is statistically significant. The best system is ARTINIT (accuracy 40.1, p-value of 0.18 and a small effect size, 0.39). Two other systems get an accuracy score better than 39, that is ADDPRED (39.3) and LININT (39.3), but not significantly better than NATONLY. The system ADDPROB, similar to ADDPRED, leads to lower accuracy, showing that adding the probabilities decrease the performance. For these systems, the scores on each of the outer folds are close[7], specially for ADDPRED,

[7] ARTINIT : standard deviation (sd) = 0.074, mean = 40.1 ; ADDPRED : sd = 0.037, ADDPROB sd = 0.061, mean $\simeq$ 39

revealing a high model stability. The other systems allow to evaluate the impact of the artificial data on the final results.

The only method leading to lower results is when training on the union of the data sets (UNION), the accuracy (22.6) is similar to ARTONLY. This was expected, as the natural data are about 372 times less numerous than the artificial ones, the new model is thus more influenced by the latter. Note that Wang et al. (2012) also experiment this setting but do not observe such a gap, maybe because their artificial data are based on manually annotated explicit examples, which are likely to be less noisy.

When directly combining the data, either by adding random subsets of the artificial data (ARTSUB, accuracy 34.5) or by weighting the natural examples (NATW, accuracy 38.9), we observe, on the inner folds, an inverse trend. As expected, the accuracy increases as the influence of the artificial data decreases, that is, decreasing the coefficients for ARTSUB and increasing the weights for NATW. Observing the results in the inner folds reveals a same trend about the relative importance of the two kinds of data: natural data have to be around 2.5 times more important than the artificial ones. We also observe this effect with LININT, with the mean of the choosen α values equals to 0.3. We also note that the variance for the values of the hyper-parameter for ARTSUB is pretty high, probably caused by the randomness of the subsamples selection. It is a bit lower for NATW and LININT showing that these methods are more robust. Nevertheless, the strategy does not give an *a priori* good value for the hyper-parameter but restricts the space of values (1020 plus or minus 272 for NATW and 0.3 plus or minus 0.18 for LININT).

6.3 Models with Automatic Selection of Examples

Previous experiments showed that adding artificial data mostly improves the performance but still not significantly. We assume that a lot of the artificial data are noisy, which could hurt the systems. The method of selection of examples thus aims at eliminating potentially noisy examples. The artificial model is used on the training set, and we keep the examples that are predicted with a probability higher than a threshold $s \in [0.3; 0.85]$ with an increment of 0.1 until 0.5 and of 0.05 until 0.85. If the model is confident enough about its prediction, the example might not correspond to noise, that is, a word form that does not have a discourse readings and/or a segmentation error. We also check whether the connective is redundant. For each threshold, we rebalance the data based on the least represented relation (+SELEC systems).

The automatic selection of examples allows to improve previous results. The accuracy of the ARTONLY model moves from 23.0 to 25.0 with selection, and the system UNION move from 22.6 to 40.1 with selection.

The best results are obtained when we use artificial data to create new features but when we add only the relation predicted by the artificial model (ADDPRED+SELEC). With this system, we observe a clear tendency toward significance (accuracy 41.7 with a large effect size, 0.756, and a high correlation, 0.842). The F_1 scores for all classes are improved : 20.8 for *contrast*, 51.0 for *result*, 31.2 for *continuation* and 53.9 for *explanation*. Two other systems get an accuracy over 40: NATW+SELEC (accuracy 41.3, with a trend toward significance[8]) and UNION+SELEC (no significantly higher than NATONLY). We note that ADDPRED corresponds to the best baseline in (Daumé III and Marcu, 2006), which shows the relevance of dealing with the distributions differences in our data through domain adaptation methods.

The automatic selection step allows a more important weight on the informations provided by the artificial data. For LININT+SELEC, the best results are obtained with an almost equal influence of the two models. In the same way, the mean of the choosen values for the coefficient for NATW+SELEC is much lower, and it increases a lot for ARTSUB+SELEC allowing for larger subsamples. Even if the choosen values are widly scattered, these observations tend to prove that the selection improves the quality of our artificial corpus. Regarding the choosen values for the thresholds, the mean over all the systems is 0.7, with a variable standard deviation but always greater than 0.1. This deviation is pretty high, this hyper-parameter probably needs a better optimisation, by repeating the inner loop for example, but these experiments will allow to reduce the search space.

[8] p-value $= 0.077$, large effect size, 0.68 and high correlation, 0.67

The automatic selection of examples leads to one system, namely ADDPRED+SELEC, significantly improving the accuracy of NATONLY. This shows that the artificial data, when rightly integrated, can thus be used to improve a system identifying implicit relations, especially if their influence is low, the model is driven towards the good distribution.

6.4 Effects on the Identification of the Relations

Looking at the F_1 score per relation, we observed that these systems have dissimilar behaviors. A larger influence of the artificial model allows improvements for *contrast*: the best result for this relation is obtained when only the artificial data are used for training (at best, 28.8 F_1 score with ARTONLY+SELEC). The identification of the relation *continuation* seems to be also improved by the influence of the artificial data. We can observe it with the linear interpolation of the models: the mean of the F_1 score increases with the increasing of the α coefficient for these relations. For *continuation*, however, the best mean F_1 is obtained with $\alpha = 0.8$, this relation needs a certain degree of influence from the natural data. Some support for this proposition can be found in the fact that the best result for this relation is obtained with NATW (at best, 44.7 F_1 score). For the other relations, a large weight on the artificial data clearly decreases the F_1 score. However, the identification of *explanation* is improved when we add the predictions of the artificial model (at best, ADDPRED+SELEC, 53.9 F_1 score). Improvement is fairly low for *result* (at best, 51.0 with ADDPRED+SELEC).

The relation *contrast* might take advantage of less noisy artificial data as most of the examples are extracted using the connective *mais* (*but*) always in discourse readings. For *explanation*, predictions of the artificial model could be quiet coherent as most of the artificial examples correspond to the pragmatic relation *explanation**. Moreover, if we look at the feature distribution (850 features overall), we observe a gap of more than 30% for 2 and 5 features for *result* and *explanation* that is not observed for *contrast* and *continuation*, the relations that make the most of the artificial data.

7 Conclusion

We have presented the first system that identifies implicit discourse relations for French. This kind of relation is difficult to identify because of the lack of specific predictors. In the previous studies on English, the performance on this task are fairly low despite the use of complex features, probably because of a lack of manually annotated data. To deal with this issue, even more crucial for French, our system also resorts to additional data, automatically annotated using discourse connectives. These new data, however, do not generalize well to natural implicit data, because of distribution differences. We thus test methods inspired by domain adaptation in order to combine natural and artificial data. We add an automatic selection of examples among the artificial data to deal with noise generated by the method of automatic annotation. We manage to get significant improvement over a system solely trained using available data manually annotated by using automatic selection and the addition of features corresponding to the predictions of the artificial model.

In future work, we will explore more sophisticated methods to deal with data samples that follow different distributions. We will also explore ways to deal with imbalanced data and use our methods on all the relations annotated in our French corpus. Finally, we will test these methods on English corpora, in order to compare their efficiency with previous studies.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus (regular paper). In *Proceedings of LREC*.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Sasha Blair-Goldensohn, Kathleen R. McKeown, and Owen C. Rambow. 2007. Building and refining rhetorical-semantic relation models. In *Proceedings of NAACL HLT*.

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010. Benchmarking of statistical dependency parsers for french. In *Proceedings of ICCL (posters)*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Joost C.F. de Winter. 2013. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC*.

Anna Gastel, Sabrina Schulze, Yannick Versley, and Erhard Hinrichs. 2011. Annotation of implicit discourse relations in the tüba-d/z treebank. *GSCL*.

David J. Hand. 2006. Classifier technology and the illusion of progress. *Statistical Science*.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue and Discourse*.

Jing Jiang. 2008. A literature survey on domain adaptation of statistical classifiers. Available from: http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.html.

Shafiq R. Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of ACL*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of EMNLP*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. Technical report, National University of Singapore.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation: Learning bounds and algorithms. In *Proceedings of COLT*.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL*.

Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of EACL*.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern Recognition*.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of SIGDIAL*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP (Short Papers)*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of ACL-IJCNLP*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. Lexconn: A french lexicon of discourse connectives. *Discours*.

Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. *Proceedings of IWPT*.

Tobias Scheffer. 1999. *Error Estimation and Model Selection*. Ph.D. thesis, Technischen Universitet Berlin, School of Computer Science.

Anders Sogaard. 2013. *Semi-supervised learning and domain adaptation in natural language processing*. Morgan & Claypool.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations, an assessment. *Natural Language Engineering*.

Caroline Sporleder. 2008. Lexical models to identify unmarked discourse relations: Does Wordnet help? *Lexical-Semantic Resources in Automated Discourse Analysis*.

Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*.

Yannick Versley. 2013. Subgraph-based classification of explicit and implicit discourse relations. In *Proceedings of IWCS*.

WenTing Wang, Jian Su, and Chew Lim Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In *Proceedings of ACL*.

Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. 2012. Implicit discourse relation recognition by selecting typical training examples. In *Proceedings of COLING (Technical Papers)*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of COLING (Posters)*.