ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-211

Preference Extraction From Negotiation Dialogues

Anaïs Cadilhac, Nicholas Asher, Farah Benamara, Vladimir Popescu and Mohamadou Seck[1]

Abstract. This paper presents an NLP-based approach to extracting preferences from negotiation dialogues. We propose a new annotation scheme to study how preferences are linguistically expressed on two different corpus genres. We then automatically extract preferences in two steps: first, we extract the set of outcomes; then, we identify how these outcomes are ordered. We finally assess the reliability of our method on each corpus genre.

1 Introduction

Information about preferences is an important part of what is communicated in dialogue. A knowledge of one's own preferences and those of other agents are crucial to decision-making [1] and strategic interactions between agents [8] [17]. Further, since agents don't come with their preferences transparently in advance, we must infer them from what an agent says or from his nonlinguistic actions, if we wish to exploit them in reasoning. Research on preferences thus divides into three subtasks [19]: extracting preferences from users, modeling of users' preferences, and reasoning about preferences to compute optimal outcomes. We focus in this paper on the first task and analyze how to infer preferences from dialogue moves in actual conversations that involve bargaining or negotiation.

A (strict) preference is commonly understood as an asymmetric, transitive ordering by an agent over outcomes, which are understood as actions that the agent can perform or goal states that are the direct result of an action of the agent. For instance, an agent's preferences may be defined over actions like *buy a new car* or by its end result like *have a new car*. The outcomes over which a preference is defined will depend on the domain or task. Among these outcomes, some are acceptable for the agent, i.e. the agent is ready to act in such a way as to realize them, and some outcomes are not. Among the acceptable outcomes, the agent will typically prefer some to others. Our goal is not to determine the most preferred outcome of an agent but rather to trace the evolution of their commitments to certain preferences as the dialogue proceeds. For example, if an agent proposes to meet on a certain day X and at a certain time Y, we learn that among the agent's acceptable outcomes is a meeting on X at Y, even if this is not his most preferred outcome. We are interested in an ordinal definition of preferences, which consists in imposing a ranking over all (relevant) possible outcomes and not a cardinal definition based on numerical values. Preference orderings can be total or partial, if some candidates are not comparable for a given agent.

We distinguish preferences from opinions. Opinions are defined as a point of view, a belief, a sentiment or a judgment that *an agent may have about an object or a person*; preferences, as we have defined them, involve an ordering on behalf of an agent and thus are *relational and comparative*. Opinions concern absolute judgments towards objects or persons (positive, negative or neutral), while preferences concern relative judgments towards actions (preferring them or not over others). The following examples illustrate this:

(a) The movie is not bad.
(b) The scenario of the first season is better than the second one.
(c) I would like to go to the cinema. Let's go and see *Madagascar 2*.

(a) expresses a direct positive opinion towards the movie but we do not know if this movie is the most preferred. (b) expresses a comparative opinion between two movies with respect to their shared features (scenarios) [15]. If actions involving these movies (e.g. seeing them) are clear in the context, such a comparative opinion will imply a preference, ordering the first season scenario over the second. Finally, (c) expresses two preferences, one depending on the other. The first is that the speaker prefers to go to the cinema over other alternative actions; the second is: given the option of going to the cinema, he wants to see Madagascar 2 over other possible movies. Reasoning about preferences is also distinct from reasoning about opinions. An agent's preferences determine an order over outcomes that predicts how the agent, if he is rational, will act. This is not true for opinions. Opinions have at best an indirect link to action: I may hate what I'm doing, but do it anyway because I prefer that outcome to any of the alternatives.

Handling preferences is not easy. First, specifying an ordering over acceptable outcomes is not trivial especially in multi criterial situations. For instance, choosing a new camera to buy may depend on several criteria (e.g. battery life, weight, etc.); hence, ordering even two outcomes (cameras) can be cognitively difficult because of the need to consider trade-offs and dependencies between the criteria. Second, users often lack complete information about preferences initially. They build a partial description of agents' preferences that typically changes over time. Indeed, users often learn about the domain, each others' preferences and even their own preferences during a decision-making process.

We are interested in how agents learn about preferences from actual conversational turns in real dialogue [13], using NLP techniques. As far as we know, this task is novel (see Section 5). Our approach to preference extraction consists of three steps, following the methodology described in [10] that builds on [2]:

1. identify dialogue segments conveying preferences and extract from each such segment the outcomes the preferences are about.
2. identify, within each relevant segment, the dependencies between the outcomes extracted at step 1 using a set of specific non-boolean operators. These dependencies allow us to infer agents' preferences and how they are ordered.
3. give a formal description of each agent's preferences. [10] proposes a procedure for translating these operators into CP-nets [7], a well-known logical formalism of representing preferences and translatable into conditional logic [20]. This provides each operator with a well-defined semantics. [10] also provides a method

[1] IRIT, CNRS and University of Toulouse, France, email: {cadilhac, asher, benamara, popescu, seck}@irit.fr

showing how CP-nets from dialogue segments combine via discourse structure to provide a model of agent preferences at any moment in the dialogue. Their model also shows the evolution of these preferences as the dialogue progresses.

This paper focuses on the first two steps of this process. We propose a new annotation scheme to study how preferences are linguistically expressed. Then, we use a machine learning approach that extracts outcome expressions from dialogues using a combination of local and discursive features. We then use a hybrid approach in order to identify the preferences over the outcomes.

2 Annotation scheme

Our data come from two corpora: one already-existing, `Verbmobil`, and one that we created, `Booking`. The first corpus is composed of 35 dialogues randomly chosen from the Verbmobil corpus [24], where two agents discuss on when and where to set up a meeting. Here is a typical fragment:

π_1 A: Shall we meet sometime in the next week?

π_2 A: What days are good for you?

π_3 B: I have some free time on almost every day except Fridays.

π_4 B: Fridays are bad.

π_5 B: In fact, I'm busy on Thursday too.

π_6 A: Next week I am out of town Tuesday, Wednesday and Thursday.

π_7 A: So perhaps Monday?

The second corpus was built from various learning resources for English, available on the Web[2]. It contains 21 randomly selected dialogues, in which one agent (the customer) calls a service to book a room, a flight, a taxi, etc. Here is a typical fragment:

π_1 A: Northwind Airways, good morning. May I help you?

π_2 B: Yes, do you have any flights to Sydney next Tuesday?

π_3 A: Yes, there's a flight at 16:45 and one at 18:00.

π_4 A: Economy, business class or first class ticket?

π_5 B: Economy, please.

We analyze how the outcomes and the dependencies between them are linguistically expressed by performing, on each corpus, a two-level annotation. First, a discourse-level annotation, splitting the text into segments (the π_i above), which are then related to each other by rhetorical relations. Second, we annotated preferences expressed by the segments. Two annotators were involved in this process.

2.1 Discourse-level annotation

Dialogues are structured by various moves that the participants make e.g., answering questions, asking follow-up questions, elaborating prior claims, and so on. Our work is novel both with respect to the literature on preference extraction and on dialogue. Existing formal models of dialogue content either do not formalize a link between utterances and preferences (e.g., [16]), or they encode such links in a typed feature structure, where desire is represented as a feature that takes conjunctions of values as arguments (e.g., [22]), making the representation too restricted to express dependencies among preferences. What is required, then, is a method for extracting partial information about preferences and the dependencies among them that are expressed in dialogue, perhaps indirectly, and a method for exploiting that partial information to identify the next optimal action.

To represent the discourse context, we use Segmented Discourse Representation Theory, SDRT [3]. SDRT structures discourse into elementary discourse units (EDUs) that are linked together with rhetorical relations such as *Question-Elaboration (Q-Elab)*, *Plan-Correction (P-Corr)*, *Question Answer Pair (QAP)*, *Plan Elaboration (P-Elab)*, *inter alia*. The segments cover a single clause of speech or complex segments composed of other segments and their relations. While the problem of extracting discourse structure remains formidable, we can approximate these relations relatively well for our purposes using features that can be conveniently obtained automatically, e.g. the presence of questions. Others like the type of discourse relations relating the current EDU to prior segments and to the EDU to come depend on an automated ability to recognize discourse relations. This is not the hardest task in discourse parsing and the prognosis is relatively optimistic [5] [25]. Our study here shows the importance of discourse features for preference extraction, assuming that these are given by manual annotation. In future work, we plan to recognize discourse structure automatically.

For `Verbmobil`, we rely on the already available discourse annotation of Baldridge and Lascarides [4]. For `Booking`, annotation was made by consensus using the same set of rhetorical relations used to annotate `Verbmobil`. To illustrate our annotation, consider again the `Verbmobil` example given in the introduction of Section 2. The corresponding discourse structures for agents A and B are respectively:

$\textit{Q-Elab}(\pi_1, \pi_2) \wedge QAP(\pi_2, \pi) \wedge \textit{P-Elab}(\pi_2, \pi)$
$\wedge\ \textit{P-Elab}(\pi_1, \pi_6) \wedge \textit{P-Elab}(\pi_1, \pi_7) \wedge \textit{P-Elab}(\pi_6, \pi_7)$,
and ,
$\textit{Q-Elab}(\pi_1, \pi_2) \wedge QAP(\pi_2, \pi) \wedge \textit{P-Elab}(\pi_2, \pi)$
where: $\pi : \textit{P-Corr}(\pi', \pi_5)$ and $\pi' : Explanation(\pi_3, \pi_4)$. Note that π and π' are complex segments composed of EDUs.

Intuitively, A's question π_1 reveals his preference for meeting next week and $\textit{Q-Elab}(\pi_1, \pi_2)$ entails that any answer to π_2 must elaborate a plan to achieve the preference revealed by π_1; this makes π_2 paraphrasable as "What days next week are good for you?", which does not add new preferences. Nevertheless, B's response in π_3 to π_5 to A's elaborating question π_2 reveals that he has adopted A's preference. In effect, A's preference is adopted in π_3, which specifies a non-empty extension for what days to meet. Inferences about B's preferences evolve as he gives his extended answer: from π_3 alone one would infer a preference for meeting any day next week other than Friday and its explanation π_4 would maintain this. But the correction π_5 compels A to revise his inferences about B's preference for meeting on Thursday. These inferences about preferences arise from both the content of B's utterances and the semantic relations that connect them together. A's response π_6 reveals that he disprefers Tuesday, Wednesday and Thursday, thereby refining the preferences that he revealed last time he spoke. A's follow-up proposal π_7 then reinforces the inference from π_6 that among Monday, Tuesday and Wednesday – the days that B prefers, A prefers Monday. This may not match his preferred day when the dialogue started: perhaps that was Friday. Further dialogue may compel agents to revise their preferences as they learn about the domain and about each other.

This example shows that agents' preferences depend upon the compositional interpretation of the discourse structure over EDUs. The constraints are different for different discourse relations, reflecting the fact that the semantics of connections between EDUs influences how their preferences relate to one another.

[2] e.g., `www.bbc.co.uk/worldservice/learningenglish`.

2.2 Preference-level annotation

To analyze how preferences are linguistically expressed in each EDU, we must: (1) identify the set O of outcomes, on which the agent's preferences are expressed, i.e. the terms, and (2) identify the dependencies between the elements of O by using a set of specific operators. Within an EDU, preferences can be expressed in different ways. They can be atomic preference statements, e.g. "I prefer X", "let's X", or "We need X", where "X" describes an outcome. "X" may be a definite noun phrase ("Monday"), a prepositional phrase ("at my office") or a verb phrase ("to meet"). They can be expressed within comparatives and/or superlatives ("a cheaper room"). Preferences can also be expressed in an indirect way using questions. That is, if A asks "can we meet next week?" he implicates a preference for meeting. For negative and wh-interrogatives, the implication is even stronger. Expressions of sentiment or politeness can also be used to indirectly introduce preferences.

Preference statements can also be complex, expressing dependencies between outcomes. We examine these negative, conjunctive, disjunctive and conditional operations over outcomes and suppose a language with non-boolean operators *not*, *&*, $\bigtriangledown$ and $\mapsto$ taking outcome expressions as arguments. A negative preference expresses an unacceptable outcome, i.e. what the agent does not prefer, as in "I am busy on Monday". As an example of conjunctive preference, consider "Could I have a breakfast and a vegetarian meal?" where the agent expresses two preferences that he wants to satisfy and he prefers to have one of them if he cannot have both. The semantics of a disjunctive preference is a free choice one. For example in "Mondays or Tuesdays are fine for me", the agent states that either Monday or Tuesday is an acceptable outcome and he is indifferent between the choice of the outcomes. Finally, some EDUs express dependencies among preferences. For example, in the sentence "What about Monday, in the afternoon?", there are two preferences: one for the day Monday, and, given the Monday preference, one for the time afternoon, at least for one syntactic reading of the utterance.

For each EDU, annotators identify how outcomes are expressed and then indicate how the preferences on these outcomes are linked using the operators *not*, *&*, $\bigtriangledown$ and $\mapsto$. In the example below, $<o>$_i indicates that o is the outcome number i in the EDU, brackets indicate how outcomes are attached and preference annotation is given after the symbol //. In π_2, annotation tells us that we have two outcomes and that the agent prefers outcome 1 over any other alternatives and given that, he does not prefer outcome 2.

π_1 : I am, going <into four>_1 or <five o'clock>_2 <on those days>_3. // 3 $\mapsto$ (1 $\bigtriangledown$ 2)

π_2 : <Tuesday the sixteenth>_1 I got class <from nine to twelve>_2? // 1 $\mapsto$ not 2

2.3 Analysis

The annotation process was performed in two steps: first a training phase where annotators jointly annotated two Verbmobil dialogues and then the dialogues of our corpora were annotated separately, discarding those two dialogues. We compute four inter-annotator agreements: (a) on outcome identification, (b) on outcome acceptance, (c) on outcome attachment and finally (d) on operator identification. We give below a brief description of our results, detailed at greater length in [9].

Using Cohen's Kappa, we obtained for (a) an exact agreement of 0.66 and a lenient agreement (i.e. there is an overlap between their text spans) of 0.85 for both corpus genres. We observed four main cases of disagreement: (1) redundant preferences which we decided not to keep in the gold standard, (2) anaphora which are often used in Verbmobil to introduce new or to precisify preferences. Hence, we decided to annotate them in the gold standard, as in "that sounds fantastic", (3) preference explanations, which we chose not to annotate in the gold standard because they are used to explain already stated preferences, and (4) finally, preferences that are not directly related to the action of fixing a date to meet but to other actions, such as having lunch, choosing a place to meet, etc. Even though those preferences were often missed by annotators, we decided to keep them, when relevant.

For (b), we got a Cohen's Kappa of 0.9 for Verbmobil and 0.95 for Booking. The main case of disagreement concerns implicit negations that are inferred from the context, as in π_1: "Tuesday is kind of out"; π_2: "same reason in the morning" where annotators sometimes failed to consider "morning" as unacceptable outcomes.

For (c), we compared how each outcome was attached to the others within the same EDU. For example, when comparing annotation $1 \mapsto (2 \bigtriangledown 3)$ with $(1 \& 2) \bigtriangledown 3$, we have three errors, one for each outcome attachment. Using F-measure, we obtained an agreement of 93% for Verbmobil and 82% for Booking.

Finally, for (d), we computed the agreements for each couple of outcomes on which annotators agreed about how they are attached. Cohen's Kappa, averaged over all the operators, was 0.93 for Verbmobil and 0.75 for Booking. We observed two main cases of disagreement: between $\bigtriangledown$ and *&*, and between *&* and $\mapsto$. In the first case, the same linguistic realization does not always lead to the same operator. For instance, in "<Monday>_1 and <Wednesday>_2 are good" we have 1 $\bigtriangledown$ 2 whereas in "I would like <a single>_1 and <a double room>_2" we have 1 & 2. In the second case, disagreements were mainly due to the difficulty for annotators to decide if preferences were dependent, or not. For example, in "I have a meeting <starting at three>_1, but I could meet <at one o'clock>_2", one annotator put *not* 1 $\mapsto$ 2 meaning that the agent is ready to meet at one o'clock because he can not meet at three, while the other annotated *not* 1 & 2 meaning that the agent is ready to meet at one o'clock independently of what it will do at three.

In the gold standard, we have a total of 1081 outcomes in Verbmobil and 275 in Booking which are located in repectively 776 and 182 EDUs. Out of these outcomes 266 and respectively 9 are unacceptable (*not* operator) in the two corpora. Finally, in Verbmobil, we have 56 instances of &, 75 of $\bigtriangledown$ and 184 of $\mapsto$ while in Booking, the counts are respectively: 31, 29 and 37.

3 Outcome extraction

Outcome extraction decides whether a given token is an outcome or not. We classify tokens into two categories: "*Outcome*" and "*Non-outcome*". Recall that outcomes can be linguistically expressed through noun phrases, prepositional phrases or verbal phrases. In the data, agents negotiate in order to reach an agreement on an action: to meet on a specified day, to book a certain flight, etc. We are generally informed about these actions in verb phrases. However, terms corresponding to preference outcomes are typically contained in noun phrases (NP). In Verbmobil, NPs denote a time or place to meet and in Booking, NPs denote specific options such as "a direct flight". Therefore, given the nature of the corpus, the presence of certain NPs and their features is a very good indicator of outcomes and the presence of preferences in EDUs. To extract NPs, we use the Charniak's syntactic parser [11].

3.1 Classifier and Feature Set

To classify NPs, we used local and discursive features, all binary. The classifier is based on Support Vector Machines. A feature vector is computed for each NP within an EDU. To assess the domain dependency of our method, we designed our features on `Verbmobil` and tested them on both `Verbmobil` and `Booking` corpora.

The scope of local features is either the NP or the EDU that contains this NP. Some of these features rely on an ontology that models a calendar (time, days, etc.) that was inspired from SUMO and COSMO[3] which are top-level-ontologies. We have five features at the NP level that test if the NP contains: a lexicalization of a concept that belongs to the ontology, a comparative, a superlative, a disjunction or a conjunction. We have ten features at the EDU level: (1) the left context of the NP is a lexicalization of a concept that belongs to our ontology. Since the list of terms associated to each concept in the ontology is small, this feature helps us to detect additional lexicalizations; the EDU contains (2) a disjunction or (3) a conjunction; scoping features that look if the NP is under the scope of (4) a negation, (5) a modal, or (6) a domain action verb (such as "*to meet*", "*to book*" and "*to reserve*"). For negation and modality, scope is resolved quite simply by using the syntactic tree of an EDU: an NP is in the scope of a negation or a modal word if the father node of that word is also a father of the NP node. Of course, this procedure does not suffice for resolving some scoping ambiguities. However, since EDUs in our data are quite short, this simple approach seems to give correct results in most cases; the EDU contains (7) an opinion word that indicates whether the agents' preferences are acceptable or not (*good*, *bad*, *OK*, etc.), (8) polite words or (9) words that introduce preferences, such as (*to prefer, to favor, favorite, choice, too, rather*, etc.); (10) the EDU restates a preference of another agent. This accounts for the "*Non-outcome*" class since it indicates that the agent does not bring new information about preferences but only repeats the already stated preferences of the other agent, as in: "you say you do not have anything open Thursday morning".

We have nine discursive features: (1-6) a set that uses the rhetorical relations that link the current EDU to the preceding or following ones. We noticed that some discourse relations can help highlight EDUs that contain preferences or not. We split discourse relations into three categories: (a) those that "generally" imply a "*Non-outcome*", such as *Explanation*, *Comment* and *Acknowledgment*, (b) those that *may* imply an "*Outcome*", such as *Elaboration*, *Continuation*, *Indirect QAP* and *Correction*, and (c) those that "generally" imply an "*Outcome*". In `Verbmobil`, 86% of discourse relations are in category (a), while 14% are in category (b). We observe the same trend for `Booking`. We thus have six features: three that test whether the relation that links the current EDU to the preceding one belongs to one of our three categories, the other three concern the relation between the current EDU and the following one; (7-8) the current EDU or the EDU that precedes it is a question. In our corpus, interrogatives are not always followed by a question mark. To detect questions, specific rhetorical relations are used, such as *QAP* and *Q-Elab*; (9) frequency feature that tests if the NP occurs at least twice in the dialogue.

3.2 Experiments and Results

Several experiments were performed for testing the validity of our extraction approach. The first experiment was carried out on `Verbmobil` (C_V). The training corpus consisted of 25 dialogues, i.e. 2374 NPs, and the test corpus consists of 10 dialogues, i.e. 700 NPs. In the second experiment (C_B), the classifier was trained on 15 dialogues from `Booking` i.e. 837 NPs and tested on 6 dialogues with a total of 312 NPs. Finally, the classifier was evaluated using `Verbmobil` for training (using the 35 dialogues) and `Booking` for test (using the 21 dialogues) (C_V + C_B). The latter, rather unusual, test configuration is supposed to help determine whether our method allows for training on a larger, already available annotated corpus and testing on smaller one, sometimes from a different domain. For all setups, we used the SVM-light software package[4].

We compared the results of the classifier with those of three baselines: (1) the first one classifies all the NPs in the "*Outcome*" category, (2) the second one classifies in the "*Outcome*" class all the NPs that contain a concept belonging to the ontology, finally (3) the third baseline is a simplified version of our classifier that only uses a subset of our features (we removed features based on ontology as well as all the features that are based on discourse relations).

Table 1 shows the results. We first present the performance of the baselines, followed by our model with only local NP features, then our model with local EDU features added, and then our model with progressive additions of discourse features (marked by the "+" sign). The last row presents the final results, obtained by using all features. The results show that, among the three baselines, the second one provides the best results for `Verbmobil`. This is expected, since the ontology is tuned to the data. However, it has limitations, because some NPs that contain a concept of the ontology are not outcomes (since they are repetitions, comments etc.) and of course not all the outcomes expressed by agents are "covered" by concepts in the ontology. For `Booking`, the ontology degrades the results (namely, the recall) with respect to the first baseline, since there is a weak overlap between the concepts in the ontology and those in this corpus. The same goes for the third test, "C_V+C_B". However, this is not a critical issue in principle, since suitable ontologies are available for the touristic domain. In all cases, the third baseline provides quite stable results, consistently better than the first baseline and, in the second and third tests (for which no suitable ontology was used) better than the second baseline as well. Interestingly, the simple classifier yields a better recall for the third test than for the second one. This might point out a data sparsity problem in training on `Booking` only (the "C_B" configuration).

The results show a similar behavior of the method on both `Verbmobil` and `Booking`. We see that the local features at the NP level are relevant for obtaining good precision. The EDU-level and the discursive features improve the recall and F-measure in all three test configurations. The improvement is more marked in the second and third tests. This might be because the ontology, less suited to these tests, has a lower impact on performance. Finally, for `Verbmobil`, we obtain an F-measure of 86.8%, i.e. almost 20% above the third baseline (simple classifier) and more than 10% above the second baseline (based on the ontology). For `Booking`, we obtain an F-measure of 64.8%, i.e. more than 10% above the simple classifier baseline. For the third test, the results do not show improvement over baselines. This is probably caused by the influence of the ontology, which better fits the support vectors to the training corpus (`Verbmobil`), making them less relevant to the test corpus. When we disable the two ontology-based features, we obtain a precision of 50.2%, a recall of 62.9% and an F-measure of 55.8%, hence, an improvement over the baselines.

[3] http://stuarthendren.net/resources/sumodlfull.owl, http://micra.com/COSMO/COSMO.owl

[4] http://svmlight.joachims.org/

Table 1. Results (in percents) for the three test configurations.

		C_V			C_B			$C_V + C_B$		
		P	R	F	P	R	F	P	R	F
Baselines	All the NP	40.9	100.0	58.1	28.0	100.0	43.8	28.3	100.0	44.1
	Ontology alone	95.6	61.3	74.7	55.6	16.7	25.7	49.2	13.5	21.2
	Simple classifier	65.2	71.1	68.0	68.4	43.3	53.1	43.9	55.7	49.1
Local Features	All features (NP)	95.7	62.0	75.2	100.0	3.3	6.5	50.7	16.0	24.4
	+ All features (EDU)	94.1	78.9	85.8	68.4	43.3	53.1	60.2	26.2	36.5
Discursive Features	+ Previous Relation	94.9	78.9	86.2	67.6	41.7	51.6	60.2	26.2	36.5
	+ Following Relation	94.0	77.5	84.9	66.7	40.0	50.0	59.4	25.3	35.5
	+ Questions	95.6	75.4	84.3	79.0	50.0	61.2	59.4	25.3	35.5
	+ $\geq$ 2 occurrences of the NP	90.8	83.1	**86.8**	75.6	56.7	**64.8**	62.9	32.9	**43.2**

As for the discursive features, we note that, for `Verbmobil`, the rhetorical relation between the current EDU and the previous one yields a more important improvement than other discourse information. This could be due to the nature of the corpus, where task context (as expressed in previous dialogue turns) is important. For `Booking`, the current EDU or the preceding one being a question yields the most salient performance improvement. This could also be due to the nature of the corpus, which mainly contains question-answer pairs at a dialogue level. For the third test, discursive features do not bring a consistent improvement over the baselines. This is perhaps caused by the inability of discourse information to compensate for the mismatch between training and test data: indeed, in principle there are more instances of local features (at the NP and EDU level) associated to positive examples, than of discursive features associated to positive examples; and when the classifier is trained on features extracted from a corpus genre and tested on another corpus genre, the weight of the discursive features might not suffice to compensate for the local features.

In all three test configurations, the feature testing for the presence of an NP at least two times in a dialogue yields consistent improvements over all other features. This is somehow expected, since NP frequency provides topicality information, and it makes sense that preferences tend to be expressed on the main topic of a discourse.

4 Preference identification

With the extracted set of outcomes from each EDU, the next step was to identify how these outcomes are ordered. To achieve this goal, we performed three subtasks: (1) first, for each EDU with more than one outcome (around 45% of the EDUs containing outcomes), we provided a structured representation of elements in O in order to get elementary couples of outcomes. For example, for π_1: "I have got a class <on Tuesday>_1 and <Thursday>_2 <from nine to twelve>_3", we get $((1,\ 2),\ 3)$; (2) the next step is to identify the set of unacceptable outcomes. This comes down to updating the structured EDU representation by adding the operator *not*. This leads to the following representation for π_1: $((1,\ 2),\ \mathit{not}\ 3)$; (3) finally, for each couple of outcomes, we recursively identify the operator that links them. For instance, for π_1 we get: $((1,\ \triangledown,\ 2),\ \mapsto,\ \mathit{not}\ 3)$. This final *EDU preference representation* is translated to a CP-net representation using a set of specific rules associated to each operator.

To perform the first subtask, we used a symbolic approach. Note that, within the structured representation, outcomes are ordered according to how their corresponding nodes are linked in the syntactic tree. In π_1, the NPs "on Tuesday" and "Thursday" are coordinated modifiers of the verb "got" and thus have a low common "grand mother" node that directly attaches to the verb whereas the NP "from nine to twelve" attaches as a separate node to the verb. We have then the tuple $((1,\ 2),\ 3)$. However, in some cases, this order has to be reversed mainly for two reasons: (1) the presence of specific discourses cues, such as "if" and "because", as in "<the twenty eighth>_1 I am free, <all day>_2, if you want to go for <a Sunday meeting>_3", where we have $(3,\ (1,\ 2))$ since the annotation should be $3 \mapsto (1 \mapsto 2)$; (2) the outcomes are not at the same ontological level, such as a day and a period of time, as in "yeah <the afternoon>_1 is okay, <on Wednesday>_2" where we have $2 \mapsto 1$. We also note that in case of some discourse cues that introduce a contrast (as "but", "although"), the syntactic order has to be modified, as in "I have class <on Monday>_1, but, <any time, after one or two>_2 I am free" where we have $(1,\ (1,\ 2))$, since the annotation should be *not* $1 \mapsto (1 \mapsto 2)$. Detecting contrasts is not easy, as relevant discourse markers are sometimes ambiguous; "but" sometimes involves contrast (see previous example) and sometimes not as in "I have a meeting, <starting at three>_1, but I could meet <at, one o'clock>_2" where we have *not* 1 & 2. The rules were built according to the same development set as for outcome extraction, i.e 25 `Verbmobil` dialogues, and were evaluated on a test set of 31 dialogues (10 from `Verbmobil` and 21 from `Booking`) that contains 412 elementary outcome couples. The F-measure is 81% for `Verbmobil` and 75% for `Booking`. These results agree well with the results we obtained on outcome attachment (see Section 2.3). Errors come both from the parser (especially for coordination attachment) and from the difficulty of detecting contrasts.

To decide whether an outcome was acceptable, we performed a binary classification task. Unacceptable outcomes are generally in the scope of lexicalized negations (*no, not*), negative opinion words (*bad*), expressions like (*I have meetings, I got classes*) or inferred from the context. Inspired by [18, 21], we designed a set of nine features: the EDU contains a negation; the outcome is in the scope of the negation; there is a delimiter between the negation word and the outcome restricting scope; the number of negation words; the number of outcomes in the EDU; the syntactic categories of the term associated to the outcome and of the negation word; the label of the negation word and finally the number of tokens between the object being classified and the negation word. We carried out a 10-fold cross-validation on both `Verbmobil` and `Booking` using a Maximun Entropy algorithm[5] with an F-measure of 89%. Observed scoping errors were due to parsing and implicit negations, as in "<Tuesday>_1 I have got a meeting <from one to three>_2 and then another one

[5] `http://nlp.stanford.edu/software/classifier.shtml`

<from four to six>_3” with 3 classified as an acceptable outcome.

The last step in our process was to identify how a couple of two outcomes from the EDU preference representation are related using the operators $\bigtriangledown$, & and $\mapsto$. As in subtask 1, we performed this step using a set of rules designed exclusively by using 25 dialogues from Verbmobil and then assessed on 31 dialogues from both Verbmobil and Booking. We got the following F-measures (results are given in the form (Verbmobil, Booking): (88%, 38%) for &, (96%, 71%) for $\bigtriangledown$, and (96%, 69%) for $\mapsto$ which correspond to an average score of 93% for Verbmobil and 59% for Booking. As for humans, our system sometimes fails to distinguish between & and $\mapsto$, between $\bigtriangledown$ and $\mapsto$, and between & and $\bigtriangledown$. Errors are more frequent for Booking because of the nature of the dialogues (more than one sentence per segment, compared to Verbmobil where segments are smaller). This makes the identification of dependencies among outcomes from distinct sentences more difficult. The errors in Booking are also due to a less clear correspondence between linguistic cues and our operators (see the discussion at the end of Section 2.3). Given that the preferences $o_1 \mapsto o_2$ and $o_1 \,\&\, o_2$ yield the same set of best outcomes, the agent is ready to act so that o_1 and o_2 are both realized. We have thus decided to collapse these two operators in order to extract, from each EDU, the preference for the best outcome. This leads to higher average F-measures of 98% for Verbmobil and 81% for Booking.

5 Related Work

We have used a linguistic approach to extract preferences from spontaneous conversation. Our linguistic approach requires both an annotation part and a preference extraction part : the annotations are needed to train and test an extraction algorithm, and the algorithm brings added value to the annotations. Our linguistic approach to preference extraction and annotation for spontaneous conversation is novel to our knowledge. In addition, no extant work uses a hybrid approach to extract preferences from dialogues using NLP techniques. Thus, there is almost no extant work for us to compare ourselves to other than [2] and [10], which we build on. Those papers show how to compute automatically preference representations for a whole stretch of dialogue from the preference representations for EDUs but do not say anything about the calculation of preferences of EDUs. Our annotation here concentrates on the preferences expressed in EDUs. We analyze how the outcomes and the dependencies between them are linguistically expressed by performing, on each corpus, a two independent annotation levels: an already existing discourse annotation [4] and, a new preference-level annotation that enhances the discourse annotation [9]. With regard to preference extraction, two main methods are used in AI to extract preferences: *preference learning* [14] where the system has to learn from users’ past preferences in order to make predictions about unseen user preferences and *preference elicitation* where preferences are the result of interactive processes with the user [12] like query learning [6], collaborative filtering [23], and more sophisticated elicitation procedures [13] [12]. While traditional preference acquisition concerns methods for getting individuals and groups to reveal preferences, these tasks don’t occur in actual conversations. So the results are not directly comparable to ours.

6 Conclusion and Future Work

In this paper, we proposed a new annotation scheme to study how outcomes are linguistically expressed on two different corpus genres. We then proposed to extract preferences in two steps: first extracting outcomes by using a machine learning approach then identifying the preferences over the outcomes using an hybrid approach. In further work, we want to test the method on larger corpora, covering various domains of conversation and also looking for outcomes in other syntactic categories (e.g VPs), to check its relevance and robustness across different domains and discourse registers. is the first step of a more complex process of preference extraction that we will completely automate in order to apply it to practical cases of negotiation or bargaining.

Acknowledgement

This work was supported by Nicholas Asher’s ERC grant number 269427.

REFERENCES

[1] N. Arora and G. M. Allenby, ‘Measuring the influence of individual preference structures in group decision making’, *Journal of Marketing Research*, **36**, 476–487, (1999).

[2] N. Asher, E. Bonzon, and A. Lascarides, ‘Extracting and modelling preferences from dialogue’, in *IPMU*, pp. 542–553, (2010).

[3] N. Asher and A. Lascarides, *Logics of Conversation*, Cambridge University Press, 2003.

[4] J. Baldridge and A. Lascarides, ‘Annotating discourse structures for robust semantic interpretation’, in *IWCS*, (2005).

[5] J. Baldridge and A. Lascarides, ‘Probabilistic head-driven parsing for discourse structure’, in *CoNLL*, (2005).

[6] A. Blum, J. Jackson, T. Sandholm, and M. Zinkevich, ‘Preference elicitation and query learning’, *Journal of Machine Learning Research*, **5**, 649–667, (2004).

[7] C. Boutilier, R.I. Brafman, C. Domshlak, H.H. Hoos, and D. Poole, ‘Cp-nets: A tool for representing and reasoning with conditional *ceteris paribus* preference statements’, *Journal of Artificial Intelligence Research*, **21**, 135–191, (2004).

[8] S. Brainov, ‘The role and the impact of preferences on multiagent interaction’, in *Proceedings of ATAL*, pp. 349–363. Springer-Verlag, (1999).

[9] A. Cadilhac, N. Asher, and F. Benamara, ‘Annotating preferences in negotiation dialogues’, in **SEM 2012*, pp. 105–113, (2012).

[10] A. Cadilhac, N. Asher, F. Benamara, and A. Lascarides, ‘Commitments to preferences in dialogue’, in *SIGDIAL*, pp. 204–215, (2011).

[11] E. Charniak, ‘A maximum-entropy-inspired parser’, in *NAACL*, pp. 132–139, (2000).

[12] L. Chen and P. Pu, ‘Survey of preference elicitation methods’, Technical report, (2004).

[13] W. Edwards and F. H. Barron, ‘Smarts and smarter: Improved simple methods for multiattribute utility measurement’, *Organizational Behavior and Human Decision Processes, 60, pp. 306-325 (1994).*

[14] *Preference Learning*, eds., J. Fürnkranz and E. Hüllermeier (Eds.), Springer, 2011.

[15] M. Ganapathibhotla and B. Liu, ‘Mining opinions in comparative sentences’, in *COLING*, pp. 241–248, (2008).

[16] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*, Oxford University Press, 2012.

[17] D. M. Hausman, ‘Revealed preference, belief, and game theory’, *Economics and Philosophy*, **16**(01), 99–115, (2000).

[18] L. Jia, C. T. Yu, and W. Meng, ‘The effect of negation on sentiment analysis and retrieval effectiveness’, in *CIKM*, pp. 1827–1830, (2009).

[19] S. Kaci, *Working with Preferences: Less Is More*, Springer, 2011.

[20] J. Lang, L. van der Torre, and E. Weydert, ‘Hidden uncertainty in the logical representation of desires’, in *IJCAI*, pp. 685–690, (2003).

[21] J. Li, G. Zhou, H. Wang, and Q. Zhu, ‘Learning the scope of negation via shallow semantic parsing’, in *COLING*, pp. 671–679, (2010).

[22] M. Poesio and D. Traum, ‘Towards an axiomatisation of dialogue acts’, in *SemDial*, (1998).

[23] X. Su and T. M. Khoshgoftaar, ‘A survey of collaborative filtering techniques’, *Advances in Artificial Intelligence*, 4:2–4:2, (2009).

[24] *Verbmobil: Foundations of Speech-to-Speech Translation*, ed., Wolfgang Wahlster, Springer, 2000.

[25] B. Wellner, J. Pustejovsky, C. Havasi, A. Rumshisky., and R. Sauri, ‘Classification of discourse coherence relations: An exploratory study using multiple knowledge sources’, in *SIGDIAL*, pp. 117–125, (2006).