Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4991–4999
Marseille, 11–16 May 2020

Inference Annotation of a Chinese Corpus for Opinion Mining

Liyun Yan[1], Danni E[1], Mei Gan[1], Cyril Grouin[1,2], Mathieu Valette[1]
[1]INaLCO, ER-TIM, 2 rue de Lille, F-75007 Paris, France
[2]Université Paris-Saclay, CNRS, LIMSI, Bât 507, Campus universitaire, F-91405 Orsay, France
liyun.yan@inalco.fr, danni.e@paris-sorbonne.fr, gan.mei@outlook.com, cyril.grouin@limsi.fr, mvalette@inalco.fr

Abstract

Polarity classification (positive, negative or neutral opinion detection) is well developed in the field of opinion mining. However, existing tools, which perform with high accuracy on short sentences and explicit expressions, have limited success interpreting narrative phrases and inference contexts. In this article, we will discuss an important aspect of opinion mining: inference. We will give our definition of inference, classify different types, provide an annotation framework and analyze the annotation results. While inferences are often studied in the field of Natural-language understanding (NLU), we propose to examine inference as it relates to opinion mining. Firstly, based on linguistic analysis, we clarify what kind of sentence contains an inference. We define five types of inference: logical inference, pragmatic inference, lexical inference, enunciative inference and discursive inference. Second, we explain our annotation framework which includes both inference detection and opinion mining. In short, this manual annotation determines whether or not a target contains an inference. If so, we then define inference type, polarity and topic. Using the results of this annotation, we observed several correlation relations which will be used to determine distinctive features for automatic inference classification in further research. We also demonstrate the results of three preliminary classification experiments.

Keywords: Inferences, Natural Language Processing, Opinion Mining

1. Introduction

The Internet provides users with the possibility to share their opinions and sentiments on various aspects of daily life. This has become a challenge for companies (reputation, customer satisfaction, etc.) and alters the way clients behave because they can freely express their negative comments whilst hiding behind their screens. These comments can be interpreted differently according to different cultural and social codes, whose diversity has a large influence on vocabulary usage (generic vs domain-specific), judgment criteria (room size, facilities and hotel services), as well as means of expression (especially elements considered negative).

There is no doubt that analysis of massive amounts of comment data requires automatic opinion mining tools to analyze the content and identify expressed trends. Further, these tools should be able to find the hidden meaning of inferences in specific contexts.

Existing tools/models can accurately predict the writer's attitude in simple explicit sentences which contain opinion words, such as 酒店地理位置很好 *(The hotel location is very good)* or 房间空间小 *(The room is small)*.[1] However, polarity is not obvious when there is inference involved. *The hotel is close to the Eiffel Tower* also means *The hotel location is very good*. In other words, the lack of opinion words in sentences may increase the difficulty of detecting the attitude of the writer. Interpreting this kind of comment requires not only textual context and domain information, but also knowledge of the cultural background of the writer. For example, 前台不讲英语 *(Receptionist cannot speak English)* is an objective statement. But for foreigners, this statement infers communication difficulty. Therefore, this objective statement is negative.

Because of this, we propose adding inference analysis to an opinion mining system. This article focuses on the annotation of inferences in Chinese texts undertaken before implementing an inference algorithm to an opinion mining system. This article is composed of:

- a detailed definition of inference, including inference types
- corpus pre-processing and a detailed explanation of our corpus annotation configuration
- statistics of annotation results
- three classification experiments using the annotated corpus
- discussion of analysis data

2. Previous Work

According to the traditional etymology, the word "inference" originates from the Latin *inferentia* which means "consequence" (Vittori, 1609). Since its appearance, application of inference has not been limited to a single domain. In basic science (McMullin, 2013), inference has been developed in geometric mathematics (Cuel, 2014), statistics (Rodriguez and Müller, 2013; Kern-Isberner and Eichhorn, 2014), linguistics (Martin, 1976; Guy and Serge, 1992), philosophy (Silins, 2013; Wright, 2014; Gjelsvik, 2015), and even Xuanzang's Buddhist teachings (Tang, 2015).

In spite of large amount of inference research, there is no consensus on neither what exactly defines an inference, nor any uniform classification of different types of inference, which are dependent on both scientific domain and on research purpose (Lavigne, 2008). In this paper, we focus on the linguistic aspects.

2.1. Definition

General dictionary and thesaurus provide the following definitions of inference: *"a conclusion reached on the basis*

[1] All the sentences in Chinese are glossed in section 3.

of evidence and reasoning" for the Oxford English Dictionary (2000) or *"a guess that you make or an opinion that you form based on the information that you have"* for the Cambridge Advanced Learner's Dictionary (Cam, 2013).

In descriptive linguistics, inference is a determining procedure for the order of meaning. A proposition *p* infers or implies a proposition *q* if and only if, *p* being true, *q* is necessarily true. Martin (2004) gives an example: if *she picked roses* is true, then *she picked flowers* is true, and considers the reasoning will no longer be valid if we reverse the two propositions. Doussau and Rigal (2011) define an inference as an operation by which we pass from one assertion considered true to another assertion based on a system of rules making the second assertion equally true. They cite a definition from the *Dictionnaire d'orthophonie* (2004) in which an inference is an additional piece of information inexplicitly contained in the message, but that the readers can deduce or assume from their own general knowledge of the world. In this way, links are established between the different parts of the text and the reader's mental representations.

Inference analysis is close to textual implication (Bedaride, 2010; Grau and Gleize, 2018). Indeed, textual implication refers to a kind of relation between the segment of source text and the segment of target text, which supposes that two texts exist both in the corpus, while the inference is rather employed in the situation that only the source text is present in the study, and it is the work of the speakers or the researchers to deduct the meaning underneath.

2.2. Inference Types

Although there have been a large number of inference studies, there is no consensus on a uniform classification of the different types of inference (Peirce, 1958; Schmalhofer et al., 2002; Lavigne, 2008; Khemlani et al., 2012), since classification is dependent on scientific field and research purpose.

Walter (1998) proposes a classification with four categories simply named by letters from A to D, which reflects the difficulty in naming inference types. Rossi and Campion (1999), van den Broek et al. (1999) and Fayol (2003) develop an inference classification which takes into account text sequencing and contains connection inference, restoration inference and elaboration inference. For example, *Mom had prepared two skirts for Julie, one red and one green. Julie did not like the red one.* The resulting interpretation is that *Julie was going to choose the green skirt*. The order of two sentences influences this interpretation. Based on the research of Peirce (1958), Dufaye (2001) advances an inference classification with induction, deduction, retroduction, in which deduction inference is distinguished by immediate inference and mediate inference (Khemlani et al., 2012). In linguistics, Duchêne (2008) distinguishes logical inference from pragmatic inference (Horn, 1984; Graesser et al., 1994; Martin, 2004; Duchêne, 2008; Vlad, 2011), while Doucy and Massoussi (2012) emphasize a distinction between lexical inference, enunciative inference and discursive inference (Patron, 2011; Ruph Porte, 2011; Ranger, 2013).

In our previous work (Yan, 2018), we defined three levels of inference analysis, based on the different inference types presented above :

- Semantic production: designates how access to meaning expressed by inference. This uses logical inference, pragmatic inference and lexical inference (defined below) (Martin, 1976; Horn, 1984; Martin, 2004; Doucy and Massoussi, 2012; Thibaud and Viviant, 2014)
- Modality production: signifies the mental process that the speaker uses to access the meaning expressed by the inference. This uses deduction, induction or retroduction (Peirce, 1958; Deledalle, 1994; Dufaye, 2001)
- Production mode: how an inference is expressed by the speaker, using enunciative and discursive inferences (Patron, 2011; Doucy and Massoussi, 2012; Ranger, 2013)

In this paper, our analysis will only be developed at the first and the last level, semantic production and production mode, which both contain logical inference, pragmatic inference, lexical inference, enunciative inference and discursive inference. We explain the differences between five types of inference with examples in the next section.

3. Corpus and Annotation Guidelines

3.1. Corpus

3.1.1. Presentation

We crawled Chinese comments of Paris hotels on two websites: booking.com and mafengwo.cn. Two different websites were selected to broaden the resource. Unlike booking.com, which is available in multiple languages for international travel reservations, mafengwo.cn is limited to the Chinese public and is solely in Chinese. The main component of the corpus consists of tourists' comments about hotels in Paris which recount their experiences and express their attitudes. The other crucial part of the corpus concerns the comment metadata, consisting of hotel score, hotel star rating, hotel location, user score, user age and user level. The comments demonstrate the specific cultural and social characteristics of these tourists in their evaluation criteria (room size, facilities and hotel service), in their choice of vocabulary (generic terms or specific terms), and in their method of expressing their sentiments (especially negative sentiments).

We collected a total of 60,000 comments. The entire corpus was segmented and tagged using Jieba, a Python Chinese segmentation module.[2]

3.1.2. Controlled Selection of Comments

Our main objective is to target the linguistic inferences in the corpus. First of all, manual annotation was needed to produce a reference corpus, mainly used for Machine Learning. Since manual annotation is both costly and time-consuming, just 1391 of the 60,000 comments were extracted for annotation,[3] accounting five percent of the cor-

[2] https://github.com/fxsjy/jieba

[3] The annotated corpus is available at https://github.com/liyunyan/ChineseHotelReviewAnnotation.git

pus. In order to increase the representativeness of the corpus extracted, we defined the following three rules to direct the extraction of sentences from the original corpus :

- Equality: The total number of positive and negative comments is equal, in order to make comparison possible.
- Superiority: The average sentence length in the annotated corpus is superior to that of the original corpus. We defined an access threshold of a minimum of eight words; this helps avoid short comments which often lack linguistic inferences.
- Balance: The percentage of the different metadata comments in the annotated corpus is positively correlated to the different metadata comments of the original corpus (e.g., the geographical distribution of comments for each arrondissement in both the annotated and the original corpus is shown in Figure 1, where the 9th arrondissement is the most over-represented in the original corpus).

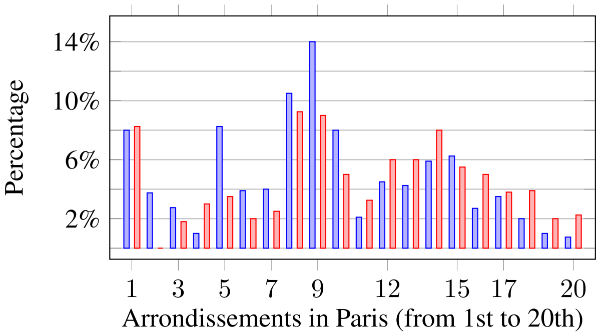


Figure 1: Distribution of comments in each arrondissement between the annotated corpus (blue) and the original corpus (red)

3.2. Annotation Framework

Our interpretation of the different kinds of inference we wished to extract was detailed in an annotation guide. The definition of each type of inference to be tagged is discussed in this section.

Annotations are made at three levels: comment, sentence, and phrase. Several kinds of information are associated with annotations made at the comment and sentence levels: (i) the presence of inference within the annotated segment (absence, presence, uncertain), (ii) the polarity conveyed (positive, negative, neutral, unknown) to be used later for opinion mining, and (iii) the topic covered by this segment (chosen from ten topics). If an inference is present, the type of inference must be specified (from five types: discursive, enunciative, lexical, logical, and pragmatic). At the phrase level, only phrases with inferences are annotated along with polarity and topic related information.

We used the BRAT annotation tool[4] to annotate our corpus since this tool allows for the rapid annotation of linguistic information in texts.

3.2.1. Annotation Levels

We consider three levels of annotation: phrase, sentence, and comment. Phrases are portions of sentences, represented by a comma, semicolon or space character, which is used like a comma in Chinese. Sentences are separated by final punctuation marks (“.” “!” “?” “...”). A comment, which is a sequence of sentences, may contain several final punctuation marks.

Our aim is to discover which inference tag is related to which level. Is there an inference type that appears more frequently at one level compared to others? We hypothesize that this three-level analysis makes it easier to distinguish inference occurrence in hotel opinion mining research.

3.2.2. Inference Annotation Tags

Presence of Inference First, the annotators should determine whether the target contains linguistic inference or not. In order to identify such inferences, the annotators focus on the topic and polarity found in the text.

- *Presence* tag: if the comment is implicit, a personal interpretation of the comment is required to obtain the binary structure: [topic, polarity], then the comment contains an inference (cf. Figure 4, 5, and 6)
- *Absence* tag: when the topic and polarity can be easily determined without any process of deduction, the comment is explicit. As a consequence, there is no inference (see Figure 2 which presents an explicit binary structure [topic, polarity], which are [location, positive] and [room, negative].

酒店	地理位置	很好。
jiǔ diàn	dì lǐ wèi zhì	hěn hǎo
hotel	location	very good

“The location of the hotel is very good”

房间	空间	小。
fáng jiān	kōng jiān	Xiǎo
room	space	small

“The room is small.”

Figure 2: Absence of inference

- *Uncertainty* tag: when a comment is unreadable (spelling mistakes), incomprehensible (use of non-standard Chinese characters), or incomplete (in Figure 3, the first sentence is not comprehensible, since we are not able to determine the topic related to the comment. The second sentence is not sufficiently complete to recognize polarity of the customer)

Inference Type If a comment contains an inference, further analysis is made to specify the type of inference. To be more precise, we distinguish five inference types:

[4] https://brat.nlplab.org/

速度	慢。
sù dù	màn
speed	slow

“The speed is slow”

*可能	是	靠近	景点
kě néng	shì	kào jìn	jǐng diǎn
maybe	be	close to	tourist attraction

“maybe (the hotel) is close to the tourist attraction”

Figure 3: Uncertainty sentences

- *Logical* inference refers to cases when the reader comes to a personal interpretation and develops his reasoning while reading the comment. In other words, this interpretation is produced by using synonymous words or grammatical information. A literal translation is enough to understand the inference and no more information needs to be added for text comprehension. For example, the sentence in Figure 6 below does not contain any opinion words, but clearly expresses the client’s negative opinion.

酒店	在	埃菲尔铁塔	旁边。
jiǔ diàn	zài	āi fēi ěr tiě tǎ	páng biān
hotel	PREP	Eiffel Tour	near by

“The hotel is located near the Eiffel Tower.”

前台	不	讲	英文。
qián tái	bù	jiǎng	yīng wén
reception	NEG	speak	English

“The receptionist doesn’t speak English.”

Figure 4: Objective statement, positive (above) and negative (below)

热 水	吹 风 机
rè shuǐ	chuī fēng jī
hot water	blow wind machine
“hot water”	“hair dryer”

Figure 5: Vocalubary with lexical inference

- *Pragmatic* inference is an inductive process which relies on knowledge acquired by an individual during past experiences. The reader brings his own world view to his interpretation of the text. The diversity of individual experience and world view means that the interpretation of the same pragmatic inference can be different depending on the reader’s experience. Because of this, more external information needs to be added to understand this kind of inference. The first example of Figure 4 contains a typical example of pragmatic inference. This objective narrative sentence implicitly expresses a positive location for accommodation. The second example is also an objective statement, since for foreigners, this statement infers communication difficulty. These instances of pragmatic inference mean that we consider the first objective statement and the second objective statement both contain negative polarity.

- *Lexical* inference is the only type which focuses on words. It presupposes that if the reader has the same knowledge as the writer, then certain words from the text contribute to inference creation. If a word is not semantically positive or negative, but it implies a positive or negative opinion towards a hotel, then it will be classified as an instance of lexical inference. For example, “hot water” is a semantically positive comment, because Chinese tourists do not drink cold tap water. Both guest comments in Figure 5 are semantically positive for Chinese tourists.

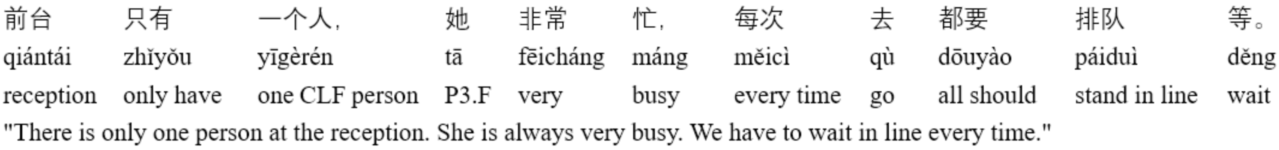
- *Enunciative* inference refers to how an inference is expressed by the speaker. It specifically occurs in three different situations. If a sentence contains an adverb that determines polarity is interrogative, or implies criticism or appreciation, the sentence includes an enunciative inference. For instance, the adverb *only* in the first sentence of Figure 7 infers an opposition of expectation. The second sentence expresses an implicit criticism.

- *Discursive* inference concerns not only one sentence, but a concatenation of several sentences. It follows that, discursive inference usually appears in long comments, especially when a topic word is not present anywhere in the comment. Due to the length of this kind of sentence and the limited length of this article, we provide a translation of the Chinese sentence: *The water in the vase of flowers hadn’t been changed for several days. We saw that the water had become turbid, which is really disgusting.*

Polarity Polarity expressed in the annotated portion is divided into four types: positive, negative, neutral (for entities that lacks an obvious polarity, such as *classical style* or *wooden structure*, etc.) or unknown (chosen only when the sentence is incomprehensible, or incomplete). Example in Figure 6 is negative.

Topic The last type of information concerns the topic covered by the comment, sentence or phrase. A total of ten topics is provided: (i) location (e.g., tourist attraction, nearby shopping, dining, environment, transport), (ii) facilities (e.g., room, bathroom, wifi), (iii) staff evaluation, (iv) cleanliness, (v) quality of service (e.g., shuttle bus), (vi) price, (vii) security (both neighborhood and inside the hotel), $(viii)$ customer base, (ix) general and (x) mixed.
Figure 8 presents an annotated example using BRAT, which is annotated in the following manner:

- The whole comment 没有电梯房间小6楼阁楼不错就是没有卫生间有卫生间的又太小转身困难基



前台	只有	一个人，	她	非常	忙，	每次	去	都要	排队	等。
qiántái	zhǐyǒu	yīgèrén	tā	fēicháng	máng	měicì	qù	dōuyào	páiduì	děng
reception	only have	one CLF person	P3.F	very	busy	every time	go	all should	stand in line	wait

"There is only one person at the reception. She is always very busy. We have to wait in line every time."

Figure 6: Logical inference

入住	三天	只有	第一天	提供	了	瓶装水。
rù zhù	sān tiān	zhǐ yǒu	dì yī tiān	tí gōng	le	píng zhuāng shuǐ
enter live	three days	only have	first day	Provide	ACC	bottle fill water

"During the three-day stay, the hotel only provided a bottle of water on the first day."

敲门	敲	了	一下	还没	来得及	应门	员工	就	进来	了。
qiāomén	qiāo	le	yī xià	hái méi	lái de jí	yìng mén	yuan gōng	Jiù	jìn lái	le
knock door	knock	ACC	once	still NEG	enough time	respond door	staff	just	enter	ACC

"(The staff) knocked on the door once. There was not enough time for us to answer before the staff entered the room. "

Figure 7: Enunciative inference

本可以房价偏贵 *(There is no elevator. There is no bathroom. In addition, the room is too small to turn around. The price is a bit expensive)* is annotated as being composed of lexical, logical, and pragmatic inferences about facilities, with a negative polarity.

- There is no sentence level in this example. Sentences are separated by final punctuation marks, while there is no final punctuation marks present in Figure 8.
- At phrase level, the annotators decide whether:
  - (a) there is no inference in the sentence 房间小 *(small room)*
  - or (b) an inference exists in the sentence 没有电梯 *(there is no elevator)*, in this last case, the inference is of lexical, logical and pragmatic types, with a negative polarity, and concerns the topic *facilities*,
  - or (c) the existence of an inference is uncertain in the sentence 转身困难 *(in addition, the room is too small to turn around).*
- At word level, the single word 电梯 *(elevator)* concerns facilities and is an instance of lexical inference with a positive polarity (nearly always).

4. Corpus Annotation

4.1. Annotation Process

The annotation work was carried out by three native-speaker Chinese annotators. Each annotation file was annotated separately by at least two annotators who then consorted together and agreed on a final version. The inter-annotator agreement scores were calculated by comparing each annotator's version with the consensus version. The writer of the annotation guide did not participate in the annotation work. The annotation corpus file was cut into 53 sub-corpus files, with 10 comments for the first 5 files, 20 comments for the next 10 files, and 30 comments for the last 38 files. The increase in the number of comments per sub-corpus file was designed to take into account the annotators' proficiency improvements throughout the task.

4.2. Inter-Annotator Agreement

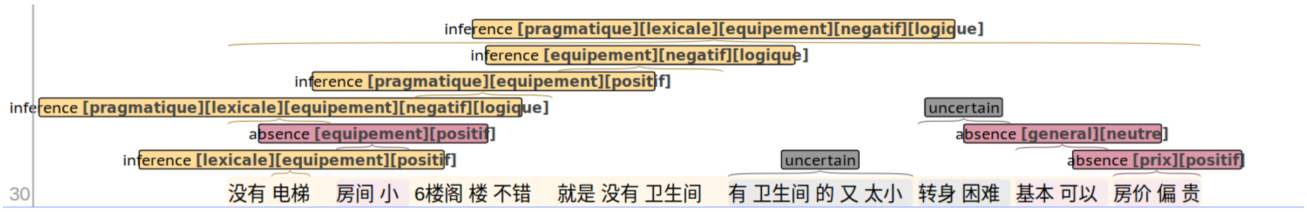
	V1	V2	V1 vs. V2
Inference presence tags	0,9627	0,9536	0,9382
Inference type tags	0,7050	0,6854	0,6505
Polarity tags	0,8449	0,8170	0,8255
Topic tags	0,7454	0,7071	0,7056
Overall	0,8145	0.7908	0,8612

Table 1: F1 score of agreement between two versions and consensus version

According to Table 1, global F1 for agreement between inference presence tags was over 0.95, which shows that each of the two annotators was in near-prefect agreement concerning inference presence tags (*presence*, *absence* and *uncertainty*), which are mutually exclusive. However, the F1 score for the inference type tags dropped to 0.6-0.7. Among the five inference type tags, we had difficulty recognizing discursive inference (with F1 at 0.3262 and 0.3642). The easiest type to identify was lexical inference. This may be because the boundary of discursive inference is more difficult to define since it often contains several sentences in one inference. As for topic and polarity tag choice, correlation of polarity tag choices was slightly higher than that of the topic tags, because topic tags had seven more candidates than polarity tags.

4.3. Quantification

Statistical results concerning our annotated corpus will be presented in this section.


inference [pragmatique][lexicale][equipement][negatif][logique]
inference [equipement][negatif][logique]
inference [pragmatique][equipement][positif]
inference [pragmatique][lexicale][equipement][negatif][logique]
uncertain
absence [equipement][positif]
absence [general][neutre]
inference [lexicale][equipement][positif]
uncertain
absence [prix][positif]
30

没有 电梯 房间 小 6楼阁 楼 不错 就是 没有 卫生间 有 卫生间 的 又 太小 转身 困难 基本 可以 房价 偏 贵

Figure 8: Annotated example using BRAT: annotations are made at comment, sentence, and word levels

	Phrase	Sentence	Comment
Presence	2740	1614	215
Absence	1637	66	0
Uncertainty	1624	76	0
Total	6001	1756	215

Table 2: Distribution of inference presence tags over three levels

4.3.1. Inference Presence Tags

For inference presence tags, we observed that the phrase level has many more inference tags than the sentence and comment levels. When it comes to the sentence and comment levels, there are fewer ambiguities, so that the number of *absence* and *uncertainty* tags are reduced, in particular at the comment level, whose number of *absence* and *uncertainty* tags were 0. In other words, all comment level reviews contain at least an inference.

4.3.2. Inference Type Tags

	Phrase	Sentence	Comment
Logical	2297	1504	210
Pragmatic	1839	1319	193
Lexical	1007	902	105
Enunciative	291	285	90
Discursive	3	96	87

Table 3: Distribution of inference type tags over three levels

In terms of the inference type tags presented in Table 3, *logical* (4012) outnumbers the other four types. The *discursive* type (187) has the least number. Furthermore, *logical* and *pragmatic* inferences more often occur at the phrase level while *discursive* occurs in the sentence and comment levels.

In addition, we also combine different inference types and calculate their frequencies. Having the most notable numbers, the three remarkable combinations are *logical+pragmatic+lexical*, *logical+pragmatic* and *lexical+logical*, while the combinations that are formed by *enunciative* or *discursive* are far fewer. This means that logical, pragmatic and lexical inferences coexist most of the time and much easier to capture, but enunciative and discursive inferences contain more information, which makes them harder to distinguish.

4.3.3. Polarity Tags

As for the number of polarity tags, unsurprisingly, the number of *positive* comments (4613) is much higher than *negative* comments(2907), with *neutral* ones appearing 390 times and *unknown* 15 times. It appears that tourists tend to make positive comments. Even negative comments contain a few moderating positive words.

4.3.4. Topic Tags

By far, the most common topic tags are *facilities* (2835) and *location* (2370), which are far more than the others. This is because tourist mostly pay attention to the facilities and location of a hotel, but also because users tend to mention various topics in one comment - this level has 699 *mixed* tags. The customer base is less important since almost all comments concern hotels with star ratings, not personal apartments, so clients have less opportunity to meet “roommates”.

5. Experimentation

5.1. Polarity Classification

Applying an emotion ontology[5] based on Ekman’s *Atlas of Emotions* to the annotated corpus, we predicted the polarity of each comment by calculating the weighting of sentiment words found in the comment. Weighting of sentiment words is in the range of -1 to 1. If a comment score is greater or equal to 1, then the comment is positive. If a score is less than or equal to -1, then the comment is negative. Scores between -1 and 1 are classified as neutral. From 7168 comments, 5069 (71%) did not match any sentiment words and 4153 (82%) of them contained at least one inference. For comments within sentiment words but whose polarity prediction were incorrect, 758 (79%) also contained an inference. We can see in Figure 9) the usefulness of inference detection in opinion mining for comments without sentiment words. Sentiment words alone without inference detection are insufficient for opinion mining.

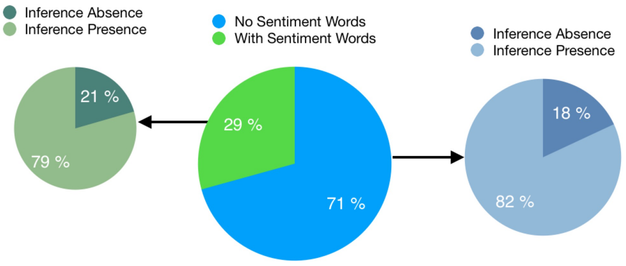


Figure 9: Proportion of inferences present in segments with and without sentiment words

[5] http://ir.dlut.edu.cn/EmotionOntologyDownload

5.2. Automatic Inference Classification with SVMs

Given that existing linguistic resources perform unsatisfactorily on opinions with inferences, we aim to automatically identify opinions with inferences, and in addition classify inference types. The modeling is based on SVMs which can efficiently perform classification on a small corpus. Our experimental features are divided into two types of metadata: hotel metadata features and morphosyntactic information. Hotel features include hotel score, star rating, user score, user age, and location. Morphosyntactic information includes comment length, negative words and number of each part-of-speech tag. There are 54 different part-of-speech tags in total. We count the number of different tags for each sentence and include them in the morphosyntactic information. Based on 7000 training data and 2500 test data, the performance of the model is presented in Table 4.

	Presence inference	Logical	Pragmatic	Lexical	Enunciative	Discursive
Accuracy	0.9185	0.8978	0.8663	0.8745	0.7560	0.9230

Table 4: Accuracy of inference classification with SVMs

5.3. Polarity Prediction with Inference

With the same training and test data, we add inference presence informations and inference types as SVM experimental features for polarity prediction. Compared to no inference modeling, using inference presence informations improves the accuracy from 0.764 to 0.9072. The performance of predicting polarity is even better by adding the five inference types with an accuracy of 0.9136.

6. Discussion

6.1. Annotation Presence and Type Tags Depending on Level

Having defined three levels of analysis, phrase, sentence and comment, we begin by analyzing the correlation of categories at different levels. It is important to know that even though we have distinguished three levels, a comment that contains only one sentence is classed as a *sentence*, not classed as a *comment*. A comment is defined as containing two or more final punctuation marks. *Sentence* is often a kind of *comment*, because *comment* represents a set of statements that ends with final punctuation. However, because of the irregular punctuation used in a web reviews, *sentence* can consist of grammatical sentences separated by either commas or spaces. This kind of *comment* is tagged as a *sentence*, since automatic classification of spaces according to function (as phrase or sentence final markers) did not obtain convincing results with an F1 of 0.409.To avoid producing noise at the analysis stage, such cases are always classified as *sentences*.

There are 6001 *phrases*, 1756 *sentences*, 215 *comments* in the annotated corpus (see Table 2). At the phrase level, each tag type is greater than 1,600. However, the numbers are greatly reduced at the sentence and comment levels, especially for *absence* and *uncertainty*, with no tags at the comment level. We can see that the presence inference tags increases as the length of tag segments decreases. In other words, inference and length of tag segments are positively correlated. This observation influenced how we locate inference during automatic inference classification. Almost no discursive inference appears at the phrase level, which demonstrates that a discursive inference involves a sequence of sentences. There are 210 logical inferences at the phrase level but many more at the sentence level (1504). This is because a logical inference is produced during the development of the text, and a phrase, rather is local or just a portion of text.

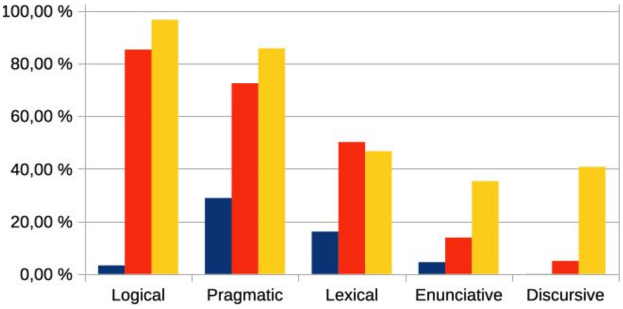


Figure 10: Distribution of five inference types for each level (*phrase* in blue, *sentence* in red, *comment* in yellow)

At the sentence level, unexpectedly, there are more discursive inferences than at the comment level. This is because the corpus contains an informal writing style where some people use spaces instead of periods as endpoints. In this kind of text, multiple grammatical sentences are classed as *sentences* rather than *comments*. This tendency also occurs in cases of lexical inference. Lexical inferences are neutral words in the general sense but carry a polarity in hotel industry discourse. In the sentence *The hotel is close to the shopping center, a large room overlooking Eiffel Tower, clean bathroom with products of l'Occitane.*, "Eiffel Tower", "shopping center" and "l'Occitane" are all lexical inferences.

In order to better understand the relationship between inference types and three levels, we converted the percentage of certain categories in each level to the Figure 10. The logical, pragmatic, enunciative and discursive types have a positive correlation from a change of phrase to comment. On the contrary, the lexical type has a negative correlation with an increase in the size of the portion.

6.2. Opinion Mining and Inference

6.2.1. Polarities and Inferences

	Positive	Negative	Neutral	Unknown
Word	1535	60	54	0
Phrase	2267	1877	225	10
Sentence	768	805	104	3
Comment	43	165	6	1
Total	4613	2907	389	14

Table 5: Distribution of polarity at different levels

According to the statistics shown in Table 5, the number of positive sentences is 30 more than negative sentences, and positive phrases 487 more than negative ones. The number of positive words is 25 times than that of negative words. At the comment, however, the number of negative comments increases abruptly up to 130 compared to 37 positive comments. We conclude that when users gave their opinions in paragraphs or with a sequence of sentences, it is very likely that they expressed a negative opinion in describing these experiences.

The *neutral* and *unknown* categories share a similar tendency. There are contained mainly at the phrase level, and appear much less often at the sentence level, and even less at the word and comment levels. In particular, there is no unknown label at the word and comment levels, because a word, except in the case of misspelled words, carries a precise meaning. It can be neutral, but never be unknown. Another for this, at the comment level, a long description contains enough information to determine the polarity.

6.2.2. Topics and Inferences

Topic	W.	Phra.	Sent.	Com.	Total
facilities	601	1619	561	54	2835
location	907	1133	310	20	2370
mixed	0	169	473	57	699
staff	48	400	124	30	602
overall evaluation	66	421	43	5	535
quality of service	7	288	98	33	426
cleanliness	0	140	23	2	165
security	10	106	29	11	156
price	0	92	18	1	111
customer base	10	8	1	1	20

Table 6: Number of topic tags at each level ordered by total number of tags (W.=Word, Phra.=Phrase, Sent.=Sentence, Com.=Comment)

We have identified ten topics with their associated subtopics in the annotation guide. As we can see in Table 6, facilities and location are the two most common topics. We can interpret this as being because facilities and location are the main criteria used when evaluating hotel stays. The topic mixed that identifies when several topics are mentioned in a single segment ranks third. The comment level has the most mixed topics, as customers usually mention more than one topic in a long description.

Among the ten topics, there are fewer opinions about the customer base. There are several reasons for this: first the hotels we analyzed have profiles on Booking and Mafengwo, unlike Airbnb, most are officially star rated hotels. Usually, guests do not share rooms or living spaces with other guests, so they rarely have an opportunity to meet other guests. The topic of the customer base is not relevant for them.

As for the word level, the *mixed* topic is unrepresented, as are *cleanliness* and *price*. That is, there is no lexical inference that concerns *cleanliness* and *price*, at least in our annotated corpus. At the comment level, the *mixed* tag is the most used, since there is more scope to talk about several topics over several sentences.

6.3. Analysis of Experimental Results

With our annotated corpus, we conducted two preliminary experiments as an example of how the corpus can be used. According to the results of our polarity classification in section 5.1, two thirds of comments do not contain any sentiment words, so that polarity cannot be simply detected by identifying words from an emotion ontology. Even if a comment does include sentiment words, polarity classification accuracy is still very low (0.5210). Therefore, automatic inference identification seems necessary in opinion mining. As for the SVM results, the model performs well, in particular for inference presence detection. Enunciative inferences are the most difficult to classify, since their average length of segments that contain them is highly variable, unlike discursive inferences (usually found in long segments), and lexical inferences (usually found in short segments). Moreover, adding inference informations as experimental features of SVM do improve sentiment polarity prediction.

6.4. Annotation Difficulties

During manual annotation, it was found that, phrase and sentence boundary recognition was not always clear-cut. Secondly, some reviews left by the tourists were syntactically incomplete but semantically comprehensible. This caused the greatest disagreement between the annotators. Also, we encountered in the comments some ambiguous words in the comments which made it difficult to determine their polarity, such as neologisms used by young people on the Internet. Even though the meaning of these words was not formally defined, they were labeled with the *lexical* inference tag, and the polarity annotated according to the intuition of annotators. All these difficulties lead to a relatively low inter-annotator agreement.

7. Conclusion and Perspectives

In this paper, we presented firstly the annotation framework we defined to process inferences. This framework was composed of three entities, five inference types, polarities and topics. From our annotation experiments, we observed that correlation relations exist within and between each level for inference types, polarities and topics. Secondly, we conducted three classification experiments to prove that inference can play an important role in opinion mining, and show that automatic identification and classification of inference types perform sufficiently well on a small text corpus. In futur research, we will begin narrowing automatic inference identification and classification by applying these regular patterns and determine the distinctive features for each type. Furthermore, our modelization will be expanded to distinguish between the phrase, sentence and comment levels.

8. Acknowledgements

We thank Jennifer Lewis-Wong for the proofreading of the paper.

9. Bibliographical References

Bedaride, P. (2010). *Textual Entailment and rewriting*. Theses, Université Henri Poincaré - Nancy 1, October.

Cambridge University Press, (2013). *Cambridge Advanced Learner's Dictionary*, 4 edition.

Cuel, L. (2014). *Discrete geometric inference*. Theses, Université de Grenoble, December.

Deledalle, G. (1994). Charles S. Peirce. les ruptures épistémiologiques et les nouveaux paradigmes. *Travaux du Centre de Recherches Sémiologiques*, 62.

Doucy, G. and Massoussi, T. (2012). Sémantique inférentielle et compréhension des verbatim clients. In *Congrès Mondial de Linguistique Française*, volume 1.

Doussau, C. and Rigal, S. (2011). Étude du développement de la production d'inférences de liaison en compréhension écrite du CE1 au CM1. Master's thesis, Université Claude Bernard Lyon1, Université Claude Bernard Lyon1, 6.

Duchêne, A. (2008). Les inférences dans la communication : cadre théorique général. In *Actes de Réducation orthophonique*. Fédération Nationale des Orthophonistes.

Dufaye, L. (2001). Les modaux et la négation en anglais contemporain. In *Cahiers de Recherche*. Ophrys.

Fayol, M. (2003). La compréhension : évaluation, difficultés et interventions. In *Actes de Conférence de Consensus*, Paris, December.

Gjelsvik, O. (2015). Rationality, capacity and inference. *Teorema: Revista Internacional de Filosofía*, 34(2):105–116.

Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text compréhension. *Psychological Review*, 101(3).

Grau, B. and Gleize, M. (2018). Textual Entailment: issues and methods for NLP. *Langages*, 4(212):105–122.

Guy, B. and Serge, B. (1992). *Compréhension de texte et sciences cognitives*. Paris, P.U.F.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. In Deborah Schiffrin, editor, *Meaning, form, and use in context: linguistic applications*, pages 10–42. Georgetown University Press.

Kern-Isberner, G. and Eichhorn, C. (2014). Structural inference from conditional knowledge bases. *Studia Logica: An International Journal for Symbolic Logic*, 102(4):751–769.

Khemlani, S., Trafton, J. G., Lotstein, M., and Johnson-Laird, P. N. (2012). A process model of immediate inferences. In *Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling*, pages 151–156.

Lavigne, J. (2008). *Les mécanismes d'inférence en lecture chez les élèves de sixième année du primaire*. Ph.D. thesis, Université Laval, Québec, Canada.

Martin, R. (1976). *Inférence, antonymie et paraphrase - Eléments pour une théorie sémantique*. Klincksieck.

Martin, R. (2004). *Comprendre la linguistique : épistémologie élémentaire d'une discipline*. Presses universitaires de France, 2e edition.

McMullin, E. (2013). The inference that makes science. *Zygon®*, 48(1):143–191, 3.

Oxford University Press, (2000). *Oxford English Dictionary*.

Patron, S. (2011). Enunciative Narratology : a French Speciality. In Greta Olson, editor, *Current Trends in Narratology*, Narratologia, pages pp. 267–289. Berlin, Walter de Gruyter.

Peirce, C. S. (1958). The collected papers of charles sanders peirce. In *Cambridge: Harvard University Press*, volume 1-6.

Ranger, G. (2013). MIND YOU : an enunciative description. In *Colloque "Modalité et commentaire / Modalisation a posteriori"*, Paris, France.

Rodriguez, A. and Müller, P. (2013). Nonparametric bayesian inference. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 9:i–110.

Rossi, J.-P. and Campion, N. (1999). Inférences et compréhension de texte. *L'Année psychologique*, 99(3):493–527.

Ruph Porte, C. (2011). Inférence lexicale et sens figuré : une entrée didactique. Master's thesis, Université Stendhal Grenoble 3, June.

Schmalhofer, F., Mcdaniel, M., and Keefe, D. (2002). A unified model for predictive and bridging inferences. *DISCOURSE PROCESSES*, 33:105–132, 03.

Silins, N. (2013). Introspection and inference. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 163(2):291–315.

Tang, M. (2015). Materials for the study of xuanzang's inference of consciousness-only (wei shi bi liang). *Wiener Zeitschrift für die Kunde Südasiens / Vienna Journal of South Asian Studies*, 56/57:143–198.

Thibaud, E. and Viviant, J. (2014). Compréhension de mots nouveaux et mécanismes d'inférences chez des enfants atteints de dysphasie âgés de 8 à 11 ans. Master's thesis, Université Claude Bernard Lyon 1.

van den Broek, P., Young, M., Tzeng, Y., and Linderholm, T., (1999). *The landscape model of reading: Inferences and the one-line construction of memory representation.*, page 71–98. 01.

Vittori, G. (1609). *Thrésor des trois langues : francoise, italiene, et espagnolle*. Genève, Philippe Albert and Alexandre Pernet.

Vlad, M. (2011). Médiation du sens et interférence dans la lecture scolaire en français langue étrangère. *Synergies Pologne*, 8:107–115.

Walter, K. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Wright, C. (2014). Comment on paul boghossian, "what is inference". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 169(1):27–37.

Yan, L. (2018). Analyse des inférences pour la fouille d'opinion en chinois. In *CORIA-TALN-RJC 2018 - Conférence sur le Traitement Automatique des Langues Naturelles*.