The Nature of Novelty Detection[*]

Le Zhao[†] Min Zhang Shaoping Ma[‡]

May 25, 2006

Abstract

Sentence level novelty detection aims at spotting sentences with novel information from an ordered sentence list. In the task, sentences appearing later in the list with no new meanings are eliminated. For the task of novelty detection, the contributions of this paper are three-fold. First, conceptually, this paper reveals the computational nature of the task currently overlooked by the Novelty community − Novelty as a combination of partial overlap (PO) and complete overlap (CO) relations between sentences. We define partial overlap between two sentences as a sharing of common facts, while complete overlap is when one sentence covers all of the meanings of the other sentence. Second, technically, a novel approach, the selected pool method is provided which follows naturally from the PO-CO computational structure. We provide formal error analysis for selected pool and methods based on this PO-CO framework. We address the question how accurate must the PO judgments be to outperform the baseline pool method. Third, experimentally, results were presented for all the three novelty datasets currently available. Results show that the selected pool is significantly better or no worse than the current methods, an indication that the term overlap criterion for the PO judgments could be adequately accurate.

Keywords: Novelty detection, overlap relations, meanings, TREC

1 Introduction

As the web gets larger and larger, there could be so many different sources of information (such as worldwide news portals, web sites of news agencies or

[*] This study was supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60223004, 60321002, 60303005, 60503064) and the Key Project of Chinese Ministry of Education (No. 104236). Part of this work was accepted by Asia Information Retrieval Symposium (AIRS 2005), as a poster "the Empirical Impact of the Nature of Novelty Detection", included in Springer LNCS series 3689/2005.

[†] State Key Lab of Intelligent Technologies and System; Dept. of Computer Science and Technology; Tsinghua University; Beijing 100084 P.R. China; zhaole@tsinghua.org.cn

[‡] State Key Lab of Intelligent Technologies and System; Dept. of Computer Science and Technology; Tsinghua University; Beijing 100084 P.R. China; {z-m, msp}@tsinghua.edu.cn

other media companies) that the user would have to browse piles of similar pages reporting the same event in seeking for some little pieces of “news”. Novelty detection is such a task that automatically removes the redundancies in the results returned by a search engine, to minimize user effort spent on finding novel information. Exactly for this purpose, the three Novelty tracks held by the Text REtrieval Conference (TREC) from 2002 to 2004 (Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004) constructed three novelty datasets as testbeds for evaluating novelty systems. The focus of the tracks was on sentence level query-specific (intra-topic) novelty detection. In the tracks, first, sentences relevant to a given topic (a query) are retrieved; secondly, according to the chronological ordering of sentences, later sentences which provide no new meanings should be removed.

Novelty detection proved helpful in information filtering (YZhang et al., 2002), personalized newsfeeds (Gabrilovich et al., 2004) and is potentially helpful for any other tasks that may return redundancies to the users.

Unlike many other natural language processing (NLP) tasks such as retrieval, summarization, machine translation or QA, which mainly deals with the relevance between documents and queries, or the syntax or meanings of documents or sentences, novelty detection is a task that deals with relations between sentences. Whether a sentence’s meanings are covered by another sentence or other sentences is its major concern, while the meanings of sentences themselves are indirectly involved. In Novelty (by Novelty, we mean the novelty detection task, and this will hold for the rest of the paper), the novelty or redundancy of a sentence is Boolean valued; sentences are either redundant because of previous sentences or novel (same as in the TREC Novelty tracks).

For Novelty, most previous works concentrated on the retrieval viewpoint of the task which saw Novelty as a single process of retrieving novel sentences from a sentence list containing possible redundancies, and thus, overlooked or neglected the nature of the novelty task we here propose − Novelty as a combination of two separate classification steps. Actually, this characteristic of the task is in a different dimension from previous works; we exploit the inter-sentence relations while previous methods focused more on finding effective features to represent natural language sentences for novelty computation (e.g. sets of terms (Zhang et al., 2002, 2003), term translations (Collins-Thompson et al., 2002), named-entities or NE patterns (Gabrilovich et al., 2004; Li and Croft, 2005), language models (YZhang et al., 2002; Allan et al., 2003; Gabrilovich et al., 2004), PCA vectors (Ru et al., 2004), contexts (Schiffman and McKeown, 2005) etc.). Since our result is independent of how individual sentences are represented, it could be applied to improve all previous novelty methods. Apart from the above possible reason, the characteristics of the available datasets could be another cause why researchers have not been able to recognize and verify the nature of Novelty we here propose (see inter-collection comparisons of Section 5); the latest collection, TREC Novelty 2004 collection, which largely supports the main result of this paper, has been available only for a short time.

In their pioneering work on Novelty, (YZhang et al., 2002) raised several fundamental questions regarding the properties of the redundancy measure and the novelty judgment procedure: symmetric or asymmetric redundancy measure, sentence-to-sentence or sentence-to-multiple-sentences comparison model. Our study has answered the two questions both theoretically and empirically.

In the TREC 2003 and 2004 Novelty tracks, there were two separate tasks. In "task 1", the participants first were required to retrieve the relevant sentences from the collection of sentences for each topic, then were required to reduce the redundancies of the retrieved sentences. In "task 2", the relevant sentences for each topic were already given, and only redundancy reductions should be performed. The focus of this paper is on task 2 − to eliminate redundant sentences and preserve all sentences that contain new information. Experimental results are also presented for the Novelty 2003, 2004 task 2 datasets and Yi Zhang et al's novelty collection (YZhang et al., 2002), in which redundancy reductions were performed on the sets of all the relevant and only the relevant sentences (documents). All the three collections are consisted of about 50 topics, with each topic a separate set of relevant sentences (documents).

TREC Novelty datasets were on a sentence level as the Novelty track organizers and participants believed that "document-level novelty detection is rarely useful because nearly every document contains something new, particularly when the domain is news" (Soboroff and Harman, 2005). Actually, the arguments of this paper are valid independent of the units for processing, and the experimental results in this work include those obtained on a document level collection constructed and used in (YZhang et al., 2002).

An outline for the rest of the paper is as follows: Section 2 is the heart of the paper, in which the two relations (PO-CO) of novelty detection computation are provided. Though our formalism can be seen as a direct derivation from the semantic theories of natural language, this general model is independent of the representations of individual sentences or documents. Implications toward techniques dealing with Novelty suggested by this computational structure will also be discussed, such as the use of language modeling and clustering techniques. Section 3 summarizes the widely used similarity, overlap and pool methods as well as current difficulties in novelty computation. In Section 4, as a direct application of the computational nature of Novelty, we try to address the current difficulties in novelty computation empirically, which leads to the selected pool method. Based on the PO-CO framework, formal error analysis for a more general family of selected-pool-like (PO-selection based) methods are presented. We provide, in Section 5, the corresponding experimental results on the three novelty detection collections, revealing the comparative advantages of the selected pool to the overlap and the pool method. Section 6 concludes the paper and proposes directions for future novelty research.

2 The two relations

Consider relations: a relation R between the elements of a set A is a subset C of the Cartesian product $A \times A$. Any $a \in A$ and $b \in A$, aRb, if $(a, b) \in C$. In this paper, A is a set of sentences, and we deal with relations between sentences.

We consider two types of relations for novelty detection: the complete overlap (CO) relation and the partial overlap (PO) relation.

2.1 CO and PO relations

First is the *complete overlap relation*. It is a partial order relation, and we denote it as $\geq_{co}$. One sentence A $\geq_{co}$ B, if A contains all the meanings of sentence B. This relation is a partial order relation. It is transitive and antisymmetric. For sentences A, B and C:

1. A $\geq_{co}$ A (Reflexivity).
2. If A $\geq_{co}$ B and B $\geq_{co}$ A, then A = B in meaning (Antisymmetry).
3. If A $\geq_{co}$ B and B $\geq_{co}$ C, then A $\geq_{co}$ C (Transitivity).

In (YZhang et al., 2002), only the third property is presented explicitly as an assumption. The above three properties together characterize the complete overlap (CO) relation.

Second, the *partial overlap relation*, which is symmetric, we denote it as $\bowtie_{po}$. A $\bowtie_{po}$ B, if A and B have meanings in common. Note that having common meanings does not require A to completely overlap B, though complete overlap is sufficient for partial overlap. This relation is non-transitive and symmetric. For sentences A, B and C:

1. A $\bowtie_{po}$ A (Reflexivity).
2. If A $\bowtie_{po}$ B then B $\bowtie_{po}$ A (Symmetry).
3. If A $\bowtie_{po}$ B and B $\bowtie_{po}$ C, A and C need not have the $\bowtie_{po}$ relation. (E.g., A = {a}, B = {a, b}, C = {b}. Here, A $\bowtie_{po}$ B and B $\bowtie_{po}$ C, but A C do not have this PO relation. No transitivity here).
4. If A $\geq_{co}$ B, $B \neq \emptyset$ then A $\bowtie_{po}$ B (Complete overlapping is sufficient for partial overlapping).
5. If A $\geq_{co}$ B and B $\bowtie_{po}$ C, then A $\bowtie_{po}$ C. Here, A is called a CO expansion of B, and this property states that CO expansions preserve PO relations.[1]
6. If A $\bowtie_{po}$ B and B $\geq_{co}$ C, A and C need not have the $\bowtie_{po}$ relation.

   The above properties (1, 2, 3, 4, 5 and 6) are the necessary conditions for the PO relation. There are also two other properties of the PO relation which are actually stronger than what is necessary:

[1] In a strict classical logic sense, $\geq_{co}$ is just implication, A implies B iff A $\geq_{co}$ B. Then, for any sentence A and B, because of the material implication of classical logic, A $\geq_{co}$ (A∨B) (disjunction of facts in A with facts in B), also, B $\geq_{co}$ (A∨B), it directly follows from this property that A $\bowtie_{po}$ B, which means any two unrelated sentences could be PO related by their disjunction. We surely do not want to see this happening. So in the strict classical logic sense, this property should be invalidated, or we could only allow conjunctions between sentences and the facts in the sentences are conjunctionally connected. In relevance logic where material implication is not allowed, this property could be allowable.

7. A $\bowtie_{po}$ B if $\exists\ C \neq \emptyset$ such that A $\geq_{co}$ C and B $\geq_{co}$ C (*Separation of meanings*). Here, C is a separated sentence containing common meanings of A and B, but need not contain all the common meanings of A and B.

8. A $\bowtie_{po}$ B if $\exists\ C = A \cap B, C \neq \emptyset$ (*The intersection definition*). Here, C contains all the common meanings.

As the PO relation is symmetric, we called the sentences that are PO related to one sentence its PO relatives. (e.g., for sentence A in {A: A $\bowtie_{po}$ B and A $\bowtie_{po}$ C}, B and C are called A's PO relatives, and similarly A is also B's and C's PO relative.)

In the above properties, (1, 2, 3, 4, 5 and 6) are the basic properties of the PO relation as they can be derived from having common-meaning definition, or from property (7) – separation of meanings alone, or property (8) alone. Property (7) is sufficiently strong for the PO relation, but may not be necessary. Property (8) is even sufficient for (7).

In the case of multiple sentences (e.g., A, B and C) overlapping one single sentence (say D), the PO relation is the case, because to have an overlap relation, A, B and C must all be D's PO relative (not necessarily $\geq_{co}$ D each), and together $A \cup B \cup C \geq_{co} D$.

Note that for sentences we assume there are also operations and relation like in set theory: $\cup$ $\cap$ and $\subset$, but they need not be exactly the same as in set theory. For example, if we need to consider novelty depending on the user (what's novel for the user, proposed by (YZhang et al., 2002)) background information and rules such as world knowledge and logic rules. The only difference it will make is that there should be corresponding modifications to the operations of the sentences, i.e. $A \cup B$ will be $A \cup B \cup$ {the facts derived from A, B and background information (if any) according to the rules}.

Although the PO relation itself is symmetric, in the novelty task, where sentences are aligned along a time line and only previous sentences can overlap a subsequent one, thus, an asymmetry is imposed onto the PO relation. We could see clearly that this asymmetry of the PO relation is external; it should not be mixed up with the other intrinsic properties of the PO relation.

Note that the relations defined above are completely different from the "partially redundant" and "absolutely redundant" in (YZhang et al., 2002), where the redundancy is more subjective, as being judged by the assessors. The two relations we here defined are more objective; they are Boolean valued, and the output of the CO relation must be either completely redundant or novel, which is closer to the notion of "absolutely redundant", thus only the "absolutely redundant" judgments in Yi Zhang et al's collection were used in the experiments of this paper.

2.2 Sets of facts

In this subsection, we provide an explanation of the PO-CO framework with semantic theories of language. Facts (which can be represented as logical expressions) are meanings of the statements that can be asserted as either true or false. In Novelty, only sentences that tell clear and complete facts are considered. Throughout the paper, we are talking about "meanings" of sentences, to be exact, it is actually the senses of sentences; we distinguish reference and sense from the ambiguous word "meaning" as (Gamut, 1991) did. Novelty requires senses, not references, because it is intensional in its nature rather than extensional, since it asks the question: "Is the sentence novel?" rather than "Is it true?". Thus, the relational structure of Novelty and the properties of the relations could be seen as having arisen from the discrepancy between the units of novelty processing (i.e. sentences) and the units of novelty definition (i.e. senses) – what is novel is actually individual senses, not sentences; sentence could be a much larger unit.

From the discussions of the previous sections, it may seem that meanings of sentences are actually treated as sets of facts, or similar to sets. We have even used sets in the examples. The set of facts assumption for meanings is strong enough to provide all previous properties listed. Since the set assumption is very strong (which may even be the strongest we can attain), we can use sets to provide counter examples, as in the PO relation property (3). But using sets introduces a problem; the set definition is too strong, and has a narrower range of application. We should generalize it little by little to the weakest assumptions we can possibly achieve.

When we define the PO relation, there are actually three different definitions. For *the first*: (A B) is a PO pair if there is common meaning between them. This definition is precise in the sense that there are no assumptions about what meanings of sentences would be like. The properties (1, 2, 3, 4, 5 and 6) of the PO relation can be derived from this definition. In spite of its simplicity, this PO definition is too ambiguous and should be formalized to bring the PO relation into the PO-CO framework. This can be achieved by *the second definition*, which defines the PO relation using the three properties of the CO relation and the *separation of meanings* (A $\bowtie_{po}$ B if $\exists\, C \neq \emptyset$ such that A $\geq_{co}$ C and B $\geq_{co}$ C). Separation of meanings is stronger than the first definition, because it says if there are common meanings, some common meanings can be separated (from A and B to a sentence C). *The third definition* is the intersection definition, which requires that for A $\bowtie_{po}$ B, there exists a maximum sentence C = $A \cap B$. Here, maximum means for any sentence S, if A $\geq_{co}$ S and B $\geq_{co}$ S then C $\geq_{co}$ S. This definition is stronger than the separation of meanings definition, since the separation of meanings can be derived from it. None of the three definitions require meanings to be treated as sets. The set assumption is even stronger than the intersection definition. That is to say, in our definitions of CO and PO relations, meanings of sentences need not be exact sets of facts.

Every assumption here has its exceptions, of course. But at least it's likely that in most cases the weak assumptions are not far from the reality or from the users' needs if we just focus our interest on the novelty detection in news stories where only simple facts and events are involved.[2] What definition we shall

[2] Applications such as abnormal state detection in industrial plant monitoring or novelty detection in robot navigation (Saunders and Gero, 2001) seem quite different from the text

choose at a specific occasion depends on what properties we need in processing, but if we adopt stronger and finer properties like separation of meanings or intersection property or even set assumption, we must be aware that the results we may attain can only be applied to more limited cases.

Here are some examples. A: "Tom has a sister; she is reading a book", B: "Tom's sister is reading a book".

Sentence A contains more meaning than B because A states that Tom has a sister (if we take only lexical information into account, implications or presumptions of sentences are not considered; B does not necessarily contain the fact that Tom has a sister). A is consisted of two facts, but B only one. So A $\geq_{co}$ B. This example can still be explained under sets of facts assumption, but is surely less obvious than in "Tom is five, and Tom goes to school" $\bowtie_{po}$ "Tom is a five-year-old boy" where the common meaning can be separated as "Tom is five". The following will be even more obscure.

C: "I frightened the cat, and it ran away", D: "I frightened the cat, so it ran away".

C contains only two facts, but D contains two facts the same as C, and also a belief that the cat ran away because I frightened it. So D $\geq_{co}$ C. If sentences become more complicated, even for the most sophisticated minds, it will be a difficult task to count the facts in them. This is especially true when we consider more background information or implications of meanings, because not only can sentences have generated meanings but also there may be contradictions derived from original sentences or the meanings implied by them. Even if the assumptions do not fail, it is still difficult to program a computer to solve them.

If, for example, we take emotional facts implied by sentences into account, it will be difficult for the separation of meanings assumption to hold: E: "You savagely killed the cat", F: "You murdered the cat". The differentiation of the emotional facts in E and F is difficult. As we consider more facets of the natural language, since the emotional subtleties of the terms like "murder" or "savagely" are hardly exact and clear, probabilistic models or fuzzy models may be of help.

Here in defining CO and PO relations, we only set up assumptions about relations between sentences, since this is the least the novelty task requires. The meanings of a sentence, whether behaving like a set or not, are not necessarily concerned. At least, we are very fortunate, as whatever definition among the three we adopt, we can always have the several basic properties of the PO relation.

As we saw in the above examples, if we are to practically use these relations, there are many factors to be defined and specified (such as what background

novelty task we are considering here; even natural language is not involved. For these applications, usually a deviancy measure for a new event to the current probabilistic model estimated from all previous scenes alone is enough. But as well, there can be similar improvements like we have brought into sentence level novelty detection: introducing a PO relation between the scenes of interest, locating the PO relatives before using the deviancy measure. This is because the PO-CO framework is a widely valid computational model for novelty detection, and exists in every novelty detection task, not necessarily text novelty detection.

information or rule to use, whether implications or presumptions of sentences are considered and setting up rules to resolve contradictions in the data, etc.), to resolve uncertainties and rule out difficult cases in the natural language.

2.3 Some direct results from the relations

After clarifying the nature of the novelty task, we can have some nontrivial examples (applications) explained under the framework of PO-CO relations.

A first example to see will be a method for Novelty that uses clustering techniques (Zhang et al., 2003) (the Subtopic III method: sentences are clustered into several classes and only sentences within one class can have an overlap relation; overlaps between clusters are not considered). As we know from the properties of the PO relation, PO relations actually differ in one point from equivalent relations: transitivity. PO relations are not transitive, thus there can be no equivalent classes. The usage of the clustering methods in the novelty task has an intrinsic difficulty - the sentences need not necessarily form classes. So introducing clustering techniques without taking this fact into account can be harmful. In TREC 2003, the Subtopic III method was shown to be ineffective. (The work (Yang et al., 2002) is different from the intra-topic clustering discussed here. In (Yang et al., 2002), inter-topic clustering of documents were performed, which is not our concern.)

Before continuing with the second example, we introduce two notions essential in Novelty: the *differentiation of meanings* and the *chronological ordering* of acquisition of knowledge; if a new sentence contains meanings that are *different* from any other *previously* known meanings (facts), it is novel. (For example, “Tom is five” is different from “Tom has a sister” even though Tom appears in both.) Consequently, if humans were unable to differentiate the meanings of sentences, the novelty task would no longer exist. And the differentiation of meanings is also sufficient for Novelty computation.

The second example is the uses of language models (LM) under this PO-CO framework. There can be two usages of LM. In the first, like in retrieval (Ponte and Croft, 1998), the generation probability of the current sentence on the language model estimated from the previous sentences can be used to estimate the probability of redundancy. Take for example, the task of ranking new documents according to their novelty (Gabrilovich et al., 2004). Given a known set of seed documents, according to the PO-CO framework, for each different newly appearing document, the LM for the previous sentences should be constructed on the PO relatives of the new document (PO-relatives for the current document in the seed collection). The document sets used to construct the language models could be different for different new documents; thus, the comparison of generation probabilities of the two new sentences using two separate LMs is not mathematically justified. This is clearer under the measure theoretic view of probability. The two different language models impose two distinct measures onto the event space (documents in the collection). In ranking documents in generation probability, we are actually measuring two objects (the two target documents) using two different rulers (the two language models). This explains the intuition that if two facts (A and B) are different and both are novel, it is impossible to judge whether A is more novel than B or not. Since Novelty requires only the *differentiation of meanings*, the ranking of documents here must have been imposed by attributes other than novelty (such as the amount of new information or the number of new meanings). In practice ((YZhang et al., 2002; Gabrilovich et al., 2004; Allan et al., 2003)), another usage of LM is common; for a current document, two LMs are constructed respectively for previous documents (as a whole) and the current document; the KL-divergence between the two models is used to approximate the degree of novelty. This use of LM, unlike generation probability, is mathematically justified. However, there is no step of finding PO relatives; all previous documents are used. Because of this, it can be easily adopted into the PO-CO framework by constructing the LM on the PO relatives.

Next, there is an important and direct implementation that benefits from the successful distinguishing of the above PO-CO relations.

2.4 Novelty – a complex task

In this subsection, we return to the novelty task itself. Once we are clear about the two relations discussed above, we can see immediately that the novelty task we used to refer to as one single task can be considered as being consisted of two separate subtasks.

The first step is to find out the pairs of sentences that share common meanings. (For a current sentence, this step is just locating the previous sentences having PO relation with the current one.) In this subtask, the *separation of meanings* definition can be useful, as in determining whether a pair has PO relation, we only need to separate some common meaning. This subtask can have its own judgment and evaluation method. We will discuss whether PO classification accuracy is necessary for the success of the PO-CO framework in Section 4.1.

The next step is to judge whether a current sentence is completely overlapped by previous PO relatives, with all the known PO pairs. We may need to separate all common meanings between two PO sentences in this subtask. And combining all the common meanings of the previous sentences with respect to the current one, we will finally be able to judge whether all meanings of the current sentence are covered by the previous sentences.

In practice, if all the sentences are short, containing only one simple fact, there is no need to use the PO relation; one CO step of one-to-one comparisons would serve the purpose well, and there will be almost no difference between the asymmetric overlap and the symmetric similarity measure. However, the longer the sentences are, the more likely multiple facts exist in a single sentence, and the more likely methods that adopt the PO-CO framework will work better than the methods that treat the complex Novelty as one single task. In the real world data, informative articles always try to include several facts in one single sentence (usually, with the help of clauses), which justifies applying the PO-CO framework to real world data.

But the two subtasks are still difficult in the sense that they have to deal with complicated cases, outliers of the simplified assumptions we proposed. Even within the scope of the assumptions, a computational solution to manipulate facts in the novelty task is still not apparent (e.g. the example sentences E and F in Section 2.2 could hardly be precisely translated into formal language sentences). But since we have broken down the novelty task into two subtasks where problems and difficulties are fewer than in the complex task that takes Novelty as a whole (also, there will be less confusions and uncertainties), it is expectable that novelty research will move a step forward.

We are now able to see that the problems mentioned in the introduction (symmetric or asymmetric novelty measure, one-to-one or multiple-to-one comparison) arose because of an unclear perception of the novelty task, and these questions are gone once we take the view from the nature of the novelty task. But this viewpoint still cannot explain how these questions arose empirically. And our study of novelty, described below, tried to investigate the empirical facet of the questions.

3 The previous methods

In the previous works, there were always two standard themes of novelty detection techniques. In one theme, to judge the current sentence, first, one-to-one redundancy comparisons between the current sentence and each of the previous sentences were performed. Next, the maximum of the redundancy scores obtained from the first step was compared against a threshold (α) to finally decide whether the current sentence is redundant; if the maximum redundancy score exceeded α, the current sentence would be classified as redundant. Simple similarity method (YZhang et al., 2002) and overlap method (Zhang et al., 2002) both adopted this one-to-one comparison paradigm. In the other theme, the redundancy score between the current sentence and the pool of all the previous sentences together was used against threshold α to make the redundancy decision. The simple pool (Zhang et al., 2002) and the interpolated aggregate smoothing Language model (Allan et al., 2003) applied this all-to-one theme.

The two themes were adopted because of the conception of the novelty detection task that in judging a current sentence, all previous sentences should be used. Later, we will show that this conception is generally wrong, because novelty judgment is not what we used to think a single inseparable judgment process.

We introduce three implementations of the above two themes related to this investigation: the similarity method (one-to-one), the simple term overlap method (one-to-one) and the simple pool method (all-to-one). Improvements over these previous methods using our approach will be shown in Section 5.

3.1 From similarity to overlap

From the *differentiation of meanings* (Section 2.3) we know that if the meaning of a sentence is the same as some known fact, the sentence is redundant. So a symmetric similarity measure between *sentences* can be used to estimate the symmetric “same” relation between *meanings*. A sentence sufficiently similar to a previous one is considered redundant. Thus, for the novelty task, it seems natural to use a similarity measure to determine whether a sentence is similar enough to a previous one. The differentiation of meanings is probably the only reasonable explanation for the use of symmetric methods in Novelty, which depends on identifying meaning (fact) with sentence; one sentence could only contain one fact.

Although similarity was proven to be effective experimentally (YZhang et al., 2002), (Zhang et al., 2003), if we think twice, when one sentence's meanings are covered by another, this relation is not necessarily symmetric, because sentences may contain multiple meanings (conjunctionally connected). An asymmetric overlap measure should be used eventually ((YZhang et al., 2002) and (Zhang et al., 2002) mentioned such belief). Actually, the overlap method in (Zhang et al., 2002) was proved to be stable among different data collections (Zhang et al., 2002, 2003; Ru et al., 2004) with a performance comparable to that of the similarity method. The similarity and overlap methods presented in this paper were defined as in (Zhang et al., 2002, 2003):

$$\begin{aligned} Sim(A,B) &= \frac{|A \cap B|}{|A \cup B|} = \frac{\sum_{i \in A \cap B} \min(A_i, B_i)}{\sum_{i \in A \cup B} \max(A_i, B_i)} \\ OverlapB_A &= \frac{|A \cap B|}{|B|} = \frac{\sum_{i \in A \cap B} \min(A_i, B_i)}{\sum_{i \in B} B_i} \end{aligned} \tag{1}$$

$OverlapB_A$ is the overlap of sentence B by a previous sentence A. A_i is the TFIDF weight of term i in A (Salton and Buckley, 1988). In experiments, thresholds were set to judge whether two sentences are sufficiently similar, or whether a large enough portion of one sentence is overlapped by another.[3]

Algorithmic description of the overlap method(Similarity is similar):

```
For the $i$th sentence in the list which may contain redundancies;
    MaxOverlap := 0;
    For $j$ := 1 to $i$-1;
        If MaxOverlap < $Overlapi_j$, Then MaxOverlap := $Overlapi_j$;
    End $j$;
    If MaxOverlap > threshold $\alpha$, Then $i$ redundant;
    Else $i$ is novel;
End $i$.
```

Surprisingly, despite the theoretical advantage of overlap, similarity is empirically better than or almost equivalent to the asymmetric methods such as the

[3] The similarity and overlap defined here are identical to the resemblance and containment measures defined in a similar context in (Broder, 1997).

overlap method, as experimental results from (YZhang et al., 2002) and (Zhang et al., 2002) indicated. (The result that asymmetric methods were worse than symmetric similarity measure of (YZhang et al., 2002) is largely invalid, because the comparison was not done on the same level, the asymmetric method used language model to represent sentences, while the symmetric method adopted a weighted term vector model.) We will present the comparisons of our overlap and similarity method in Section 5.

3.2 The pool method

In the above subsection, only sentence to sentence comparison is considered. But for general novelty detection, since all "old" sentences should be used to judge the current sentence, a method that compares the current sentence with all previous sentences would be more justified. A pool method would be an obvious choice, in which overlap between the pool of terms from all previous sentences and the set of terms of the current sentence is computed, with a fixed threshold α for redundancy judgment like in overlap.

Algorithmic description of the simple pool method:

```
P := $\emptyset$; //Initializing the pool P.
For $i$ := the 1st to the last sentence in the list;
    If $Overlapi_P >$ threshold $\alpha$, Then $i$ redundant;
    Else $i$ is novel;
    P += Sentence $i$;//updating pool with all the terms in $i$
End $i$.
```

But features like TFIDF weighted terms, being only surface features of sentences not the exact meanings, make this pool consisting of terms from all previous sentences too noisy to perform well. If all the terms of a target sentence appeared in its precedences, it will be classified as redundant, however, simple term appearance does not mean that the same meaning of the term has already appeared, even if the meaning of a term is the same as in a previous sentence, the fact that the term was used to express could still be different ("Tom has a sister, she is five" completely overlaps "Tom is five" in terms, but the two sentences differ in meaning). Thus, on one hand, simple pool has a large false redundancy rate, on the other hand, simple overlap could return more redundant sentences as it excludes multiple-to-one overlap cases. A selected pool method based on the nature of the task could resolve this difficulty.

4 The selected pool method

From the stance of the PO-CO framework, it is clear that the previous overlap and pool methods came about because of the ambiguous conception of the novelty task that when making *the novelty judgment* of the current sentence, we could and should use all the previous sentences in the list, while as a matter of fact, all the previous sentences should be used in the PO judgment, not necessarily the CO step. Accordingly, we propose a selected pool method, in which only sentences that are related to the current sentence are included in the pool (the PO step), followed by a pool-sentence overlap judgment (the CO step). In the experiments, if the TFIDF overlap score of the current sentence by a previous sentence exceeded the selection threshold β, that previous sentence was considered to be PO related to the current sentence. By setting the threshold β to be 0, we include all previous sentences in the pool - the selected pool turns back into the simple pool method. Setting β to be the threshold for pool-sentence overlap judgment α, the selected pool becomes the simple overlap method. Table 2 shows the change in the performance of the selected pool method as β changes. The relative performance of the selected pool and pool methods compared to the baseline overlap method, with automatically learned parameters α and β using cross validation will be provided in the next section, which is summarized as Table 3. What we must point out is that the term overlap score as a feature for making PO and CO decisions is very coarse and certainly not perfect, which suggests possible further improvements. The selected pool method solves the dilemma faced by the simple pool and overlap methods, as we could avoid the noisy simple pool with a selection step while at the same time consider multiple-to-one overlap cases among sentences, which cannot be achieved using the simple overlap method.

Algorithmic description of the selected pool method:

```
For the $i$th sentence in the list;
    S := $\emptyset$; //Initializing the selected pool S for $i$.
    For $j$ := 1 to $i$-1;
        If $Overlapi_j > \beta$, Then S += $j$; //$j$ selected as $i$'s PO-relative.
    End $j$;
    If $Overlapi_S >$ threshold $\alpha$, Then $i$ redundant;
    Else $i$ is novel;
End $i$.
```

There is another thing about using the PO-CO framework in Novelty computation that is worth noting. Following the two relations in Novelty, the final and best unit for processing novelty would seem to be facts (i.e. logical expressions formalized from sentences). Unfortunately, without precise and sufficient formalizations that could satisfy the retrieval needs of every user, computers could hardly use facts correctly for computation. And the task of translating natural language sentences into logical forms is still far too difficult. When we do novelty detection, we have to use units such as documents or sentences to base our computation on. Therefore, the two classification steps (PO: the step of classifying whether two sentences are PO related, and CO: classification of whether PO relatives of a sentence $\geq_{co}$ the current sentence) always exist. (Here we use the word classification in the sense as in Pattern Classification, by Duda et al. (2000)). Even if NLP systems were able to extract exact meanings of sentences (disregarding whether this is generally possible, problems like the examples E and F of Section 2.2 brought up), in that case, without adopting the PO-CO framework, the simple pool method would be precise enough for Novelty computation. However, since the pool method needs to maintain an increasing pool of all the occurred facts from the collection, it is memory consuming. Thus, the selected pool method we here introduced would still be a useful option for the system designer to choose, demanding only a minimum amount of memory − the size of the target sentence.[4] This shows that the PO-CO framework in novelty is a general structure, which exists independent of how sentences are expressed. Therefore the framework could be applied to tasks other than text novelty detection, just as passage 4 of Section 2.2 pointed out.

4.1 Error analysis of the selected pool

For measuring the performance of the methods based on this PO-CO framework like that of the selected pool, there is always a question: how good must the PO relative selection be to attain a better performance than the pool method, which assumes all previous sentences to be PO related to the target sentence[5]. The analysis in this subsection shows that the inexactness of the representation of natural language sentences would yield the pool method inferior to the selected pool. This analysis will show why the pool method is "noisy" empirically, as discussed in Section 3.

Consider the strict pool and selected pool methods where one new word would yield the current target sentence novel. We take it for granted that only the terms from the sentences are accessible to the selected pool and pool methods, and used to infer the meanings of the terms. Taking into account the polysemy phenomena of natural language, the meanings of the same term could differ in different sentences. Below we provide a rigorous analysis of how the strict selected pool method could be better than a strict pool method.

Definition 4.1 *A **sentence** A is a set of unique terms $A = \{a_i | \forall i \in \{1...n\}$ and $\forall j \in \{1...n\}$, $a_i \neq a_j$ if $i \neq j\}$. $|A|$ is the number of terms in A.*

Definition 4.2 *The **meaning of a sentence** A − $Mean(A)$ is defined to be the collection of meanings of its individual terms $\{a_i^A | i \in \{1...n\}\}$. For different terms, a_i in sentence A and b_j in sentence B (if B is just A, a_i and b_j are from the same sentence), a_i^A may equal b_j^B (Synonyms).*

[4] (Opitz et al., 1999) discovered precisely such a case where the PO-CO framework was adopted because of a limited working memory in the human brain. (Opitz et al., 1999) identified two steps in the human brain during novelty processing – the "*retrieval of related semantic concepts*" at the right prefrontal cortex (which usually actively maintains context information during performance of working memory tasks) and "*registration of deviancy*" at the superior temporal gyrus (the language and music processing center); the two steps corresponds exactly to the PO and CO relations we revealed.

[5] As the overlap method excludes the cases multiple sentences overlapping a target sentence, it is generally erroneous. Because of this, although sometimes the overlap method could perform better than the pool method, we still only provide error analysis of a simplified version of the selected pool method against the pool method.

Remark (Polysemies)[6]: Meanings of terms are context dependent. If a_i^A is the meaning of term a_i in sentence A, for any other sentence B, also containing term a_i, a_i^A does not necessarily equal to a_i^B.

Remark: The above definition of meanings is just a simplification of the true meanings of natural language sentences (not exactly sets of facts as the examples in Section 2.2 showed). This definition captures the polysemy and synonym phenomena of the natural language.

Definition 4.3 *For a sentence C with sentences $\{A_i|i \in \{1...m\}\}$ preceding it, a **novelty measure:** Nov (Nov could be pool, selected pool or any other measure) is a classifier $Nov(C|A_i)$ which returns 0 if C is judged redundant and 1 if novel.*

Definition 4.4 *The **selection method** of selected pool is a classifier $sel(A, B)$ which infers whether a sentence A is PO-related to the sentence B, $sel(A, B) = 0$ if judged not PO-related and $sel(A, B) = 1$ if judged as related.*

Remark: Though the PO relation is symmetric by definition (Section 2), a selection method could be asymmetric. A selection method could only infer the meanings of sentences from the observed terms, while the PO relation is by definition a relation between sentences in their meanings.

Definition 4.5 *A novelty measure Nov is said to be **strict** if: $\forall$ target sentence C and $\{A_i|i = 1...n\}$ preceding it, $Nov(C|\{A_i|i = 1...n\}) \equiv 1$ iff $\exists j \in \{1...|C|\}$, $\forall i \in \{1...n\}$, $\forall k \in \{1...|A_i|\}$, $c_j \neq a_{ik}$ holds.*

Remark: Since only terms are accessible to the novelty measure (meanings can only be inferred), a strict novelty measure is a measure that would judge the target sentence novel for one new term appearing in the sentence. We consider only strict novelty measures in our analysis because we need to fix the CO method when comparing the PO selection methods and a strict measure is both simple and direct to serve as this baseline CO method.

Definition 4.6 *For the current target sentence C and its only precedence A (with $A \bigcap C = \{a_i; i \in \{1...|A \bigcap C|\}\}$, the common terms of A and C), a selection method is said to be **adequately accurate** if*
$[1 - P(sel(A, C)|\forall i, a_i^A = a_i^C)] \cdot P(\forall i, a_i^A = a_i^C) <$
$[1 - P(sel(A, C)|\exists i, a_i^A \neq a_i^C)] \cdot P(\exists i, a_i^A \neq a_i^C)$.[7]

Consider the two cases, case 1 (correct exclusions): excluding a previous sentence that does share polysemies with the current sentence would yield the final novelty judgment of the current sentence to be novel, which is correct, as

[6] Since synonyms will not differentiate the pool and the selected pool methods (both handle synonyms as unrelated words), we could exclude the synonym phenomena from our analysis, and only need to consider the effects of polysemies.

[7] Here, $P(sel(A, C)|X)$ is the probability for a particular selection method sel to select sentence A for target C, conditioned on event X.

the meanings of certain terms in the current sentence are different from that in the previous sentence; case 2 (selection misses): failure in selecting a previous sentence that does not share polysemies with the current sentence (which means term meanings of the previous sentence are the same as in the current sentence) would make the selected pool method worse than the simple pool. The adequately accurate condition for a particular selection method means that the probability of the first case happening should be larger than the probability of the second case.

Theorem 4.1 *Consider strict novelty measures, with A the only sentence preceding the target sentence C, the adequately accurate condition for the selection method is necessary and sufficient for a selected pool method to outperform the pool method in novelty classification accuracy.*[8]

Proof. Without loss of generality, we assume, for the terms of A and C, $C \subseteq A$ and $c_i = a_i, \forall i \in \{1...|C|\}$. (For, if C contained a new term, the strict pool and selected pool would both render it novel, and there would be no difference between the two measures.) Consider the only two cases:

1. C is novel, i.e. $P(novel) = P(\exists i, a_i^A \neq a_i^C)$. The probability for the strict pool method to yield a correct novelty decision is $P(pool|novel) = 0$, since it always classifies C to be redundant. For selected pool, the same correctness rate is $P(selpool|novel) = 1 - P(sel(A, C)|\exists i, a_i^A \neq a_i^C)$.

2. C is redundant, $P(redundant) = P(\forall i, a_i^A = a_i^C)$. $P(pool|redundant) = 1$, $P(selpool|redundant) = 1 - P(sel(A, C)|\forall i, a_i^A = a_i^C)$.

Immediately we have,
$P(pool) < P(selpool)$
$\Leftrightarrow P(novel) \cdot P(pool|novel) + P(redundant) \cdot P(pool|redundant) < P(novel) \cdot P(selpool|novel) + P(redundant) \cdot P(selpool|redundant)$
$\Leftrightarrow$ the selection method is adequately accurate. QED

Corollary 4.2 *With a perfect selection method the strict selected pool is always no worse than the pool method.*

Proof. For a perfect selection method, $P(sel(A, C)|\forall i, a_i^A = a_i^C)$ always equals to 1, but $P(sel(A, C)|\exists i, a_i^A \neq a_i^C)$ could be less than 1. Therefore, a perfect selection method is always adequately accurate. QED

One observation from this adequately accurate condition is that the better performance of a selected pool method is guaranteed through the existence of polysemies. The more probable the same term in different sentences have different meanings, the more probable the selected pool method wins. As in many realistic cases, this probability of polysemy could be low, the ratio $\frac{1-P(sel(A,C)|\exists i, a_i^A \neq a_i^C)}{1-P(sel(A,C)|\forall i, a_i^A = a_i^C)}$ need to be much larger than 1 to make selected pool win. Thus, we need $P(sel(A, C)|\forall i, a_i^A = a_i^C)$ to be almost 100% and

[8] For more than one sentence preceding the target sentence, there could be a similar definition of adequately accurate and a similar theorem of necessary and sufficient condition. We present only the case for one previous sentence for clarity of presentation.

$P(sel(A,C)|\exists i, a_i^A \neq a_i^C)$ to be as low as possible to satisfy the adequately accurate condition. This could be achieved by using a conservative selection method that tends to classify sentences to be PO related. For the selected pool methods in Section 5, a selected pool with a lower selection threshold β yields more PO relatives for any target sentence and thus is a more conservative selection method. Conforming to this analysis, experiments in the next section show that with lower selection thresholds, the selected pool is always better than the pool method, while for larger thresholds it is not always the case (see Figure 5.3 and Table 2 for details). This empirical success of using term overlap as a selection criteria indicated that term overlap could be an adequately accurate selection measure.

Another observation is that although a perfect selection method is always no worse than without a selection method (as in the pool method), a better selection accuracy is not a guarantee of a better novelty classification accuracy. Thus, when trying to improve the PO method, we should use the overall novelty classification efficiency as the final evaluation criterion, rather than the PO classification accuracy alone.

5 Experiments and analyses

5.1 Novelty data sets

We start this section by introducing the novelty collections used in this work.

In Yi Zhang et al's pioneering work on large scale empirical study of the novelty detection problem (YZhang et al., 2002), a document level novelty detection dataset (nvyiz) was constructed on the archive of news articles from Associated Press (AP) year 1988 to 1990 and Wall Street Journal (WSJ) 1988 to 1990. This collection has totally 50 topics, but 5 of them lacks human redundancy assessments which were excluded from the experiments in this paper. In (YZhang et al., 2002), two notions of redundancy were used in the assessments: absolutely redundant and somewhat redundant. In the experiments below, we are concerned only with the notion of absolute redundancy which is the same as from the TREC Novelty collections.

In TREC 2003 and 2004, two datasets TREC Novelty 2003 (nv03) and 2004 (nv04) were constructed, also on newswire articles. Both consist of 50 topics, but use sentences as units of processing instead of documents (sentences from 25 relevant articles for each topic were used to construct nv03 and nv04). The TREC Novelty 2002 collection contained too few redundancies (23 of the 50 topics had ALL relevant sentences marked as novel) (Harman, 2002), thus was excluded from the experiments.

Experiments in this paper were performed on these three collections, the only public text collections currently available for novelty detection research.

Table 1: Similarity and the overlap method

nv04 5 docs	#ret	Av.P	Av.R	Av.F	#novel
s0.4	986	0.688	0.977	0.790	627
o0.7	974	0.694	0.964	0.786	634
nv04 25docs	#ret	Av.P	Av.R	Av.F	#novel
s0.4	7008	0.463	0.957	0.610	3282
o0.7	6965	0.462	0.950	0.608	3255
nv03	#ret	Av.P	Av.R	Av.F	#novel
s0.4	13495	0.719	0.978	0.817	9962
o0.7	13303	0.719	0.972	0.815	9836
nvyiz	#ret	Av.P	Av.R	Av.F	#novel
s0.4	9082	0.919	0.977	0.946	8313
o0.8	9349	0.909	0.988	0.945	8452

5.2 Similarity and overlap

Table 1 provides for each run: #ret - the total number of sentences for the 50 topics returned by a run (judged to be novel by a run), Av.P - precision of the true novel sentences in the returned averaged over 50 topics, Av.R - average recall of novel sentences, Av.F - average F-measure (F-measure trades off between precision and recall), and #novel - number of novel sentences returned. In all the tables, we used the following abbreviations: "s α" for similarity with threshold α, "o α" for overlap with threshold α, and "p α" for pool. In Table 1, "o0.7" is the overlap method with threshold $\alpha = 0.7$; "s0.4" is the similarity method with $\alpha = 0.4$. (both overlap and similarity thresholds were chosen to be optimal on the test collection.)

From the table, we can see that in F-measure, similarity was slightly better than or almost equivalent to overlap on the three collections: nv03 (0.817 vs. 0.815, but not significant by the sign-test), nv04 (for all the 25 documents, 0.610 vs. 0.607 but not significant by the sign-test, for the first 5 documents, 0.790 vs. 0.786, not significant; we keep the results from the first 5 documents for nv04 due to the reason that later sentences in the collection could admit more PO-relatives, and could turn out to be an unfair comparison, since overlap generally returns less novel sentences than similarity), and nvyiz (0.946 vs. 0.945, similarity is better, but not significant). The F-measure difference was small because overlap and similarity differ slightly. However, this comparison showed that asymmetric measures could work as efficient as symmetric measures for novelty, quite contrary to (YZhang et al., 2002)'s observation.

Table 2: Overlap and pool as special cases of the selected pool method with a varying parameter: β

nv04	β	#ret	Av.P	Av.R	Av.F
p0.7	0.0	5713	0.495	0.864	0.615
sp0.7	0.2	6068	0.490	0.900	0.621
sp0.7	0.3	6331	0.483	0.918	0.618
sp0.7	0.4	6624	0.473	0.931	0.613
sp0.7	0.5	6818	0.466	0.942	0.609
o0.7	0.7	6965	0.462	0.950	0.608
nv03	β	#ret	Av.P	Av.R	Av.F
p0.7	0.0	9127	0.755	0.762	0.744
sp0.7	0.6	13250	0.720	0.969	0.815
o0.7	0.7	13303	0.719	0.972	0.815
nvyiz	β	#ret	Av.P	Av.R	Av.F
sp0.8	0.5	8981	0.824	0.866	0.844
sp0.8	0.6	9180	0.914	0.974	0.942
sp0.8	0.7	9257	0.912	0.984	0.945
sp0.8	0.75	9296	0.911	0.985	0.946
o0.8	0.8	9349	0.909	0.988	0.945

5.3 The selected pool method

We provide experiments comparing the selected pool to the overlap method[9], and the advantage of the selected pool method to the simple pool method, on Novelty 2003 (nv03), 2004 (nv04) and Yi Zhang et al's collection (nvyiz). Analyses concerning the different characteristics of the three collections and the differences in the relative performance of the discussed methods are also present.

In Table 2, we provide the performance change as parameter β changes. In the following discussion, we use "spα sβ" as an abbreviation for the selected pool method with CO threshold α and PO selection threshold β. As β changes from 0.0 to α, the selected pool method (spα sβ) changes gradually from the pool (spα s0.0) to the overlap method (spα sα). The selected pool with a higher selection threshold will include fewer sentences in the pool, and thus will return more sentences than with a lower selection threshold. Thus we could use the number of the additional returned novel sentences in the totality of the extra returned sentences to measure performance change. For the nv04 collection, in F-measure sp0.7s0.2 is better than p0.7, and sp0.7s0.5 is almost the same as o0.7. But for the additional returned sentences, only a small portion were novel (for the 355 more sentences returned by sp0.7s0.2 than p0.7, only 147 were novel; for the 147 more returned by o0.7 than sp0.7s0.5, only 37 were novel), much lower than the average precision of about 0.49. Simple derivation showed that to increase

[9] Although overlap has a theoretical disadvantage for excluding multiple-to-one cases in novelty judgments, it has a stable and high empirical performance. Therefore, we will use the overlap method as a baseline in this experimental section.

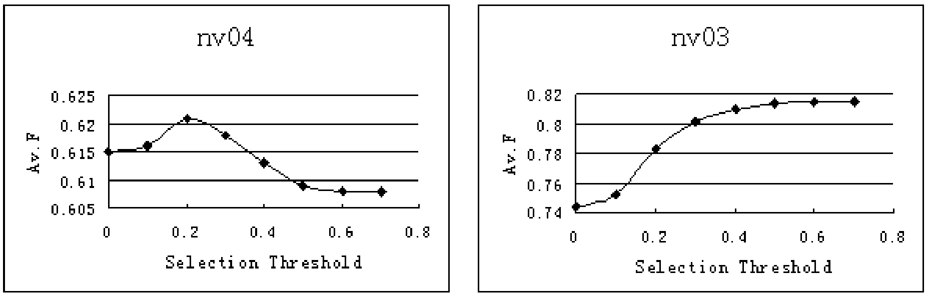

Figure 1: The performance of selected pool as selection threshold β changes

the F-measure of a set of results, additionally returning a set with precision higher than P÷(P+R) is sufficient, where P and R are the precision and recall of the original result set. For example, if P=0.5 and R=0.9, including a set with precision greater than 0.36 already increases F-measure. This property of the F-measure can be misleading when comparing different Novelty methods only using the F-measure (this could also be observed in Table 1 when comparing the similarity and the overlap method).

The performance outline of the selected pool method as selection threshold β changes is shown in Figure 5.3 (the plot of nvyiz is similar to that of nv03 and thus excluded from the figures). Consistent with the expected performance of a successful selection method as the analysis in Section 4.1 revealed, selected pool on both nv03 and nv04 collections experienced performance improvements from the pool method ($\beta = 0$) for low selection thresholds ($0 < \beta < 0.35$). This is because the term overlap selection method for low threshold β keeps PO recall high ($P(sel(A,C)|\forall i, a_i^A = a_i^C$) large) while maintaining a low false alarm level ($P(sel(A,C)|\exists i, a_i^A \neq a_i^C$) small), such that it satisfies the adequate accurate condition. This indicates that term overlap being a highly effective feature for novelty computation is also a good enough feature for using the selected pool method.

5.3.1 Within-collection analyses

In the comparisons below, a robust statistical test (Hull, 1993) − the sign test was used.

On nv04, the F-measure for sp0.7s0.2 (the best performing selected pool) was significantly better than that for o0.7 (the baseline overlap method of the best selected pool) by the sign test, significant at p = 0.0002, of all the 50 topics, 37 increased, 11 decreased, 1 remained the same; if we consider #errors made in novelty judgments, the improvement of sp0.7s0.2 w.r.t o0.7 is more conspicuous (44 topics decreased in #errors − improved, 5 increased − degraded, 1 remained unchanged). On the nvyiz collection, sp0.8s0.75 was almost the same as o0.8 in average F-measure (0.946 vs. 0.945, p=0.30, not significant by the sign test. But in #errors, 15 topics improved, 7 degraded, 23 did not change, sp0.8s0.75 was significantly better than o0.8 at p=0.037). On the nv03 collection, sp0.7s0.6 was equivalent to o0.7. These experiments on the three collections suggested that multiple-to-one comparison is no worse and sometimes better than one-to-one comparison if we use a proper method like the selected pool.

Now we are able to answer the question mentioned in the introduction, one-to-one or multiple-to-one comparison, empirically. In (YZhang et al., 2002), the multiple-to-one comparison was actually an all-to-one comparison, like in the simple pool method, and simple pool was significantly worse than overlap on nvyiz and nv03 (0.744 vs. 0.815, significant at p=0.0000000001) collections, but was significantly better than overlap on nv04 (by the sign test, significant at p=0.05), suggesting that the pool method is unstable among datasets.

For the nv04 collection, the best performance of the selected pool (sp0.7s0.2) was observed around the top runs submitted to TREC 2004 task 2 (Best run: City U. Dublin average F-measure: 0.622, second: Meiji F: 0.619), among all runs with language modeling approaches, cosine similarity measure, information gain and named entity recognition (Soboroff, 2004). Even the worst (simple overlap) could be ranked as high as 7^{th}. The overall performance of the selected pool method as a technique that adopts the PO-CO framework is encouraging.

5.3.2 Inter-collection experiments and analyses

Above are analyses within collections; inter-collection comparisons of the collections themselves and of the performance differences of the methods on different collections are provided below:

Although nv03 and nv04 are datasets selected from the same set of topics and from the same newswire data collection, the redundancy rate by human assessments in nv03 is 34.1% while in nv04 53.7%. This difference was surprising but unexplained in (Soboroff, 2004). We believe that this difference in human assessed redundancy ratio is the cause for the difference in performance of selected pool on the nv03 and nv04 collections. Selected pool is better on nv04 than nv03 probably because of the possible different characteristics in the Novelty 2003 and 2004 human assessments - 04 contains more multiple-to-one overlapping cases while one-to-one dominates in 03. There is no direct evidence for this conjecture (human assessments are incomplete for nv03 and nv04 collections; we do not know a sentence is redundant because of which previous sentences), but it seems to be the most probable explanation for the different behaviors of selected pool. It is possible that with a much shorter list of relevant sentences for each topic, (the rate of relevant sentences is almost less than half that of nv03, this allows assessors to consider multiple-to-one overlap cases more easily) when they were constructing the nv04 dataset, the assessors paid more attention to the multiple-to-one overlapping cases. This is consistent with the observation from Figure 5.3 on nv04, where the performance of selected pool dropped from the peak for larger β while on nv03 there was no such decrease as β became larger. This could also be a feasible explanation for the higher redundancy percentage in nv04 than that of nv03 which actually were consisted of topics randomly chosen from the same collection.

Compared to the nv03 and nv04 collections, nvyiz has a more complete structure (for each redundant document, the human assessments also include all the previous documents that actually make this document redundant). Therefore we can have direct evidence from the human assessments showing that nvyiz has a per topic redundancy rate of 10.8%, multiple-to-one cases occupy about 34.7% of the 10.8% redundancies. The existence of those multiple-to-one cases indicates a potential of improvement for the selected pool over the simple overlap method.

One last thing about the comparison between overlap and selected pool is how to choose the parameters α and β. As selected pool has one degree more freedom than overlap − parameter β, does selected pool tend to overfit because of its superior learning ability? To answer this question, we did Leave-One-Out (LOO) estimations to estimate the expected F-measure and expected #errors of overlap and selected pool. In these experiments, for each topic, the other 49 topics were used for training; the one topic left out was used for validation. Because the parameters were few, the entire parameter space was searched at the training step. Sign tests on F-measures of the 50 (45 for nvyiz) test topics showed that on nvyiz selected pool was better than overlap in F-measure (0.946 vs. 0.945, but not significant, p = 0.30); on nv04 selected pool was significantly better than overlap (0.621 vs. 0.614, significant at p=0.036[10]); on nv03 selected pool and overlap performed almost the same (0.815 vs. 0.815, overlap was slightly better, but not significant). The important thing here is that the performance of selected pool estimated by LOO is almost the same as the performance of the selected pool with the best parameter setting, which means the selected pool is stable and does not overfit training data, in spite of its greater learning ability (containing one more free parameter than that of the overlap or the simple pool).

The performance of the selected pool and the simple pool method compared to the overlap on the three collections (nv03 nv04 and nvyiz) are summarized in Table 3[11]. In the table, "−−" stands for significantly worse than overlap on the corresponding collection; "++" stands for significantly better under both F-measure and #errors than the overlap method; "+" stands for improvement in average F, but not significant; "0" stands for almost no difference.

At this point, we are able to answer the questions proposed by (YZhang et al., 2002) both theoretically and empirically. The poor performance of the asymmetric language model approach of (YZhang et al., 2002) was because of the particular usage of language model in that work. If we use sets of terms to represent sentences in novelty computation, the symmetric similarity method is not significantly better than the asymmetric overlap. The worse performance of multiple-to-one comparison theme is because of the failure of the authors to

[10] Since we only made pairwise comparison of selected pool to overlap, there was no need to make multiple comparison corrections.

[11] We compared the selected pool with the overlap, and compared the simple pool with the overlap separately. We show that the simple pool is unstable among collections, while the improvement of the selected pool is consistent. Since only pairwise comparisons were made, multiple comparison corrections were not necessary.

Table 3: Performance of the pool and the selected pool method compared to the overlap method respectively

Collections:	**nv03**	**nv04**	**nvyiz**
Simple pool	−−	++	−−
Selected pool	0	++	+

recognize the computational structure − the two relations of novelty detection computation. In fact, the multiple-to-one method examined in this paper, the selected pool, is better than or no worse than the baseline simple overlap method which was proved stable previously.

6 Conclusions and future work

The major contribution of this paper is the recognition of the PO-CO relations of novelty computation and the clarifying discussions and analyses (such as the *differentiation of meanings*, the classification viewpoint and the error analysis of PO-CO based methods). The nature of novelty detection we revealed is important because it provided new insights to the novelty task theoretically and empirically, and provided more flexibility for a computational solution to the novelty task.

To be more specific, previous works only adopted the one-to-one or all-to-one themes in novelty processing, however, a PO-relatives-to-one theme become possible after the recognition of the nature of the task. Previous works used uniform representations of sentences in the two PO-CO steps, while generally, since the PO and CO relations are largely independent, we could use distinctly different representations or methods for the two steps. To find out ways to exploit this flexibility for improving novelty detection accuracy would be an important topic for future novelty research.

As the two steps constitute one novelty classifier, if we only want to achieve a better novelty detection accuracy (measured by the F-measure or number of classification errors), the two steps become inter-related. An effective PO-relative classifier is only effective relative to its next CO step. However, if the PO classification is 100% accurate, it is guaranteed that an all-to-one novelty method will be no better.

The empirical success of the selected pool method proved the effectiveness of using term overlap for PO and CO judgments (if time is ripe, NLP techniques would be more desirable here). Our analysis also revealed the necessary and sufficient condition for the PO-selection based methods to outperform the pool method in the existence of polysemies. We hope to find more general situations where PO-selection based methods could work better. For the CO step, we restricted our analysis to the strict case. We hope to loosen this constraint in our analysis in the future.

For incorporating background knowledge of the user (personalized novelty detection), the PO-CO framework could also provide us a more efficient alternative than the simple pool method which could include even more noises into the pool.

Although the experiments in this study were on query-specific novelty detection datasets, many of the conclusions obtained in this paper can have a larger generalization to non query-specific cases. For the novelty task itself, there is still much work to do following this direction, but we hope this work as a summary for one major aspect of the three years' work on Novelty can be a starting point for those who would like to continue the quest for efficient novelty computation. For the study of semantics, the novelty task also provides us a new insight into the characteristics of meanings: the overlapping relations between meanings of sentences. In our treatment, unlike previous theories which consider meanings themselves, we studied meanings of sentences with the relations between them. Although these treatments are far from complete, they do work for novelty computation at least. Nevertheless, novelty detection remains a difficult task which is demanded by the complexities and arbitrariness of natural language. In this study, we identified where exactly the difficulties lie, and divided the them into small pieces. We hope this new treatment could inspire new methods and insights to tasks dealing with complicated objects such as the meanings of natural languages.

7 Acknowledgements

Special thanks to Ellen Voorhees and the anonymous reviewers of Journal of Information Retrieval for numerous suggestions about the organization, the presentation, the evaluation methodologies of this paper, and several important issues related to the PO relation property (5) and the semantic modeling of sentences in Section 4.1 that has made this paper greatly better. The authors would also like to thank Prof. Xiaotie Deng for help on the final revision of the paper that improved its readability a lot.

References

Allan, J., Wade, C., and Bolivar, A. (2003), "Retrieval and novelty detection at the sentence level," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2003)*, pp. 314–321.

Broder, A. (1997), "On the resemblance and containment of documents sequences," in *Compression and Complexity of Sequences 1997*, pp. 21–29.

Collins-Thompson, K., Ogilvie, P., Zhang, Y., and Callan, J. (2002), "Information filtering, novelty detection and named-page finding," in *Proceedings of the eleventh Text REtrieval Conference (TREC 2002).*

Duda, R., Hart, P., and Stork, D. (2000), *Pattern Classification, 2nd Ed*, Wiley-Interscience.

Gabrilovich, E., Dumais, S., and Horvitz, E. (2004), "Newsjunkie: Providing personalized newsfeeds via analysis of information novelty," in *Proceedings of the 13th international conference on World Wide Web (WWW 2004)*, pp. 482–490.

Gamut, L. (1991), *Logic, Language and Meaning*, Chicago: the University of Chicago Press.

Harman, D. (2002), "Overview of the TREC 2002 novelty track," in *Proceedings of the eleventh Text REtrieval Conference (TREC 2002).*

Hull, D. (1993), "Using statistical testing in the evaluation of retrieval experiments," in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1993)*, pp. 329–338.

Li, X. and Croft, W. (2005), "Novelty detection based on sentence level patterns," in *Proceedings of ACM Fourteenth Conference on Information and Knowledge Management (CIKM 2005)*, pp. 744–751.

Opitz, B., Mecklinger, A., Friederici, A., and von Cramon, D. (1999), "The functional neuroanatomy of novelty processing: Integrating ERP and fMRI results," *Cerebral Cortex*, 9, 379–391.

Ponte, J. and Croft, W. (1998), "A language modeling approach to information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1998)*, pp. 275–281.

Ru, L., Zhao, L., Zhang, M., and Ma, S. (2004), "Improved Feature Selection and Redundance Computing - THUIR at TREC 2004 Novelty Track," in *Proceedings of the 13th Text REtrieval Conference (TREC 2004).*

Salton, G. and Buckley, C. (1988), "Term weighting approaches in automatic text retrieval," *Information Processing and Management*, 24, 513–523.

Saunders, R. and Gero, J. (2001), "Designing for interest and novelty, motivating design agents," in *Proceedings of the ninth international conference on Computer aided architectural design futures*, pp. 725–738.

Schiffman, B. and McKeown, K. (2005), "Context and learning in novelty detection," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005).*

Soboroff, I. (2004), "Overview of the TREC 2004 Novelty Track," in *Proceedings of the 13th Text REtrieval Conference (TREC 2004).*

Soboroff, I. and Harman, D. (2003), "Overview of the TREC 2003 Novelty Track," in *Proceedings of the twelfth Text REtrieval Conference (TREC 2003).*

— (2005), “Novelty Detection: The TREC Experience,” in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005).*

Yang, Y., Zhang, J., Carbonell, J., and Jin, C. (2002), “Topic-conditioned novelty detection,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (SIGKDD 2002).*

YZhang, Y., Callan, J., and Minka, T. (2002), “Novelty and redundancy detection in adaptive filtering,” in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2002)*, pp. 81–88.

Zhang, M., Lin, C., Liu, Y., Zhao, L., and Ma, S. (2003), “THUIR at TREC 2003: Novelty, robust and web,” in *Proceedings of the twelfth Text REtrieval Conference (TREC 2003)*, pp. 556–567.

Zhang, M., Song, R., Lin, C., Jiang, Z., Jin, Y., Liu, Y., Zhao, L., and Ma, S. (2002), “Expansion-based technologies in finding relevant and new information: THU TREC2002 novelty track experiments,” in *Proceedings of the eleventh Text REtrieval Conference (TREC 2002).*