Mining Opinion Holders and Opinion Patterns in US Financial Statements

Chien-Liang Chen
Dept. of Computer Science
National Chengchi University
Taipei, Taiwan
g9813@cs.nccu.edu.tw

Chao-Lin Liu
Dept. of Computer Science
National Chengchi University
Taipei, Taiwan
chaolin@nccu.edu.tw

Yuan-Chen Chang
Dept. of Finance
National Chengchi University
Taipei, Taiwan
ycchang@nccu.edu.tw

Hsiang-Ping Tsai
Dept. of Finance,
Yuan Ze University.
Taoyuan, Taiwan.
hptsai@saturn.yzu.edu.tw

Abstract—**Subjective statements provide qualitative evaluation of the financial status of the reporting corporations, in addition to the quantitative information released in US 10-K filings. Both qualitative and quantitative appraisals are crucial for quality financial decisions. To extract such opinioned statements from the reports, we built tagging models based on the conditional random field (CRF) techniques, considering a variety of combinations of linguistic factors including morphology, orthography, predicate-argument structure, syntax and simple semantics. The CRF models showed reasonable effectiveness to find opinion holders in experiments when we adopted the popular MPQA corpus for training and testing. We also identified opinion patterns in the form of multi-word expressions (MWEs), which is a major contribution of our work. In a recent article published in a prestigious journal in Finance, single words, rather than MWEs, were reported to indicate positive and negative judgments in financial statements.**

Keywords: semantic labeling, opinion mining, financial text mining, sentiment analysis, information extraction, conditional random fields

I. INTRODUCTION

Opinion mining and sentiment analysis have been widely discussed in not only Computer Science but also Finance. Loughran and McDonald [14] developed positive and negative unigram lists that better reflect the tones of the U.S. financial statements (10-K filings) than the words in traditional psychology dictionaries, and they examined the linkage between the textual statements and the financial figures.

We propose a computational procedure to model the text in financial statements, and employ conditional random field (CRF) models for opinion holder identification and subjective opinion patterns extraction. We used a variety of linguistic features including morphology, orthography, predicate-argument structure, syntax and simple semantics. For effectively tuning and evaluating the CRF models, we trained and tested the models with the annotated MPQA corpus [17]. The goals of our work include identifying opinion holders and extracting subjective opinion patterns which are in the form of multi-word expressions. We employed the best performing CRF model that we found in a sequence of experiments to identify opinion holders and extract subjective opinion patterns in U.S. financial statements.

The major contribution of our work is to find a way to automatically expand the lists of words that are linked to positive and negative financial statuses into subjective multi-word expressions (MWEs). Opinion patterns include opinion holders and subjective multi-word expressions (MWEs). For instance, the opinion patterns in the sentence "*The Company believes the profits could be adversely affected*" include opinion holder "*The Company*" and two subjective expressions: "*believe*" and "*could be adversely affected*". Unlike traditional models that considered individual words (unigrams) and "bag of words" [16], our methods attempt to automatically extract MWEs from textual statements. MWEs could capture the subjective evaluations of the financial statuses of the reporting corporations more precisely, so are potentially more informative than individual words to facilitate better decision makings of creditors and investors. In recent years, researchers implemented quantitative methods to investigate the relationship between financial performance and the text contents of financial press. Antweiler and Frank studied the influence of Internet stock message board on the stock markets by using both 1.5 millions messages posted on Yahoo! Finance and Raging Bull that talked about 45 companies in the Dow Jones Industrial Average. A naïve-Bayes algorithm was applied to measure bullishness in messages. The results concluded that the stock messages could help predict market volatility but not stock returns [1]. Li relied on the information in the texts of annual financial statements to examine the implications of risk sentiment of corporation's 10-K filings for stock returns and future earnings. Risk sentiment of annual reports is measured by the frequencies of the individual words that are related to the risk or uncertainty in 10-K filings. Li found that the risk sentiment is negatively correlated with future earnings and future stock returns which can be predicted by the risk sentiment under cross-sectional situation [11]. Tetlock et al. concluded that negative words contained in financial press about firms included in the S&P 500 capture some hard-to-be-quantified firms' fundamentals, and they can forecast low earnings well. Although the markets underreacted to negative words in firm-specific news, the results showed that the negative words, especially those related to the fundamentals of corporations, are useful predicators for both earnings and returns [22].

Liu applied techniques for opinion mining and sentiment analysis in Computer Science to find the text or speech that carry opinions, sentiment or emotion [13]. Text in financial statements roughly belongs to two types: fact discourse or opinion discourse. The former provides objective information about the specific object (e.g., entities, events or topics). In contrast, the latter contains subjective evaluation, belief, judgment or comments about the objects. The main difference between them is whether or not subjective expressions are included in the discourses. Assuming that human's subjective feelings about objects have only two extremes, i.e., pleasantness and unpleasantness, we categorize opinion expressions into three kinds of polarities: positive, negative and neutral. For a given polarity, human beings may have different degrees of feelings about different things. For example, "wrong" and "might be inaccurate" are negative words expressing disapproval of someone or something, but the word "wrong" conveys stronger disapproval than "might be inaccurate"

978-0-7695-4601-8/11 $26.00 © 2011 IEEE
DOI 10.1109/TAAI.2011.19

pragmatically.

Researchers employed different machine learning techniques to determine the sentiment in text of different granularities. Some worked at the document-level; others may work on the paragraph-level, the sentence-level, the phrase-level or the word-level. Pang et al. utilized the concept of naïve Bayes, maximum entropy classification and support vector machines to classify the sentiment polarity of movie reviews at the document-level [18]. Weibie et al. classified subjective sentences based on syntactic features such as syntactic categories of the constituent [23]. Riloff and Wiebe used a bootstrapping machine learning process to extract the subjective expressions from sentences, manually annotated both the strength and polarity of the subjective expression, and collected them into the subjective word list [20]. Kim and Hovy used syntactic features to indentify the opinion holders in the MPQA corpus by a ranking algorithm that considered maximum entropy [8]. Choi et al. adopted a hybrid approach that combined conditional random field (CRF) and the AutoSlog information extraction learning algorithm for identifying sources of opinions in the MPQA corpus [3].

The paper is organized as following. Section II briefly introduces the U.S. financial statements and the MPQA corpus; Section III elaborates linguistic features and the CRF models that we used to mine the opinion holders; Section IV and V, respectively, discusses experimental results when we used the MPQA corpus and the U.S. financial statements as the sources of information; and Section VI concludes the paper.

II. FINANCIAL STATEMENTS AND ANNOTATED CORPUS

Financial statements are the SEC 10-K filings of public companies which download from the EDGAR database [5] of the U.S. Securities and Exchange Commission. Since the data volume of all 10-K filings is too huge to process, we selected 2,102 sample filings from the population of all filings for further processing. Instead of elaborating the processes of sentence selection and subjective opinion patterns extraction that would be discussed in Section V, we describe the annotated corpus which is engaging in training and testing our supervised CRF model in the remaining paragraphs.

Recently, many opinion mining works used MPQA (Multi-Perspective Question Answering) corpus for semantic annotation labels [17]. MPQA corpus covers different topics of news from different news sources. Since the annotation unit of MPQA is one sentence per data, we focus on the sentence-level opinion holder identification and opinion patterns extraction.

For training opinion holder identification model, we used MPQA corpus to get the opinion holder identification model and also selected part of annotation types as the tagging labels which included five different aspects of labels "*agent*", "*expressive-subjectivity*", "*objective speech event*", "*direct-subjective*" and "*target*". For better identifying opinion holders, the IOB format is used which has widely employed in NP chunking, word segmentation and named entity recognition research. In Table 1 "according to" would be tagged as "B-*objective speech event*" and "I-*objective speech event*", which "B" stands for beginning word of phrase and "I" stands for continue word of phrase; the single word "believe" in "*direct-subjective*" label which is both beginning and continue word would be tagged as "B-*direct-subjective*". The labels are not overlapped and mutually embedded. We defined the opinion holders as a phrase with label "agent" in the corpus while the expression is implicit or explicit. Instead of limiting opinion holder to be a person, it can be any entity that expresses opinion, belief, speculation and private state to object directly or indirectly. The detail example is included in Table 1.

Table 1 A sample sentence and annotations of the MPQA

According to Datanalisis' November poll, 57.4 percent of those polled feel as bad or worse than in the past and 55.3 percent believe their main problems are in the economic area.	
MPQA annotation labels	
Opinion holder 1	Datanalisis' November poll: agent; According to: objective speech event.
Opinion holder 2	57.4 percent of those polled: agent; feel: direct-subjective; as bad or worse: expressive-subjectivity.
Opinion holder 3	55.3 percent: agent; believe: direct-subjective; main problems: expressive-subjectivity.

III. CRF MODELS AND FEATURE SETS

The main task of the paper, opinion holder identification, is viewed as a sequential tagging problem which uses features in morphology, orthography, predicate-argument structure, syntax and simple semantics to train the linear-chain conditional random field (referring Lafferty et al. [10] for using linear-chain CRF model in the sequential tagging problem). Although the relationship between surface manifestations and semantic role labeling is still indeterminate, the linking theory proponents argue for prediction of semantic role labeling by syntactic information and predicate-argument structure is feasible [7]. Because the opinion holder identification is a sub-domain of semantic role labeling research, it is feasible and reasonable to use general linguistic features for opinion holder identification.

Since the data is divided into sentence unit, the granularity of the feature set which includes token-level, phrase-level, and sentence-level features but no cross-sentence features. The Figure 2 in appendix shows our linear CRF data view, we use example sentence "*We decided to make some bold decisions*" and partially selected feature set which is consist of "f1, f2, f7, f8, f9, f10, f11, f12 and f17" to clarify how the features of token-level, phrase-level and sentence-level are transformed into linear data view. The token-level feature value of each token in a sentence is extracted and recorded sequentially, but sentence-level feature value of a sentence is processed only once. In addition, the detail interpretation of phrase-level features that extracted from syntactic parse tree is explained at syntactic category feature (f11) and illustrated by syntactic parse tree in Figure 1.

The task of feature values processing was completed by Stanford NLP toolkits [21], ASSERT semantic role labeler [2][19] and CGI Shallow parser [8] for linguistic features. The following is the linguistic feature set being taken for CRF model.

A. Morphological and Orthographical features

Original token (**f1**): we separated the words in the sentence by both the white space and punctuations and also kept its original form without further being processed, so the single English words, numbers, symbols and punctuation are viewed as one token respectively except for taking named entity with periods or numbers as a single token.

Lemmatization (**f2**): the tokens above may contain many syntactic derivation and pragmatic variations. Since the different part of speech of words derived from the same lemma word would be semantically equivalent, and the lemma word usage can reduce the complexity of feature spaces.

We used regular expression to recognize the following orthographical features which might indicate the named entities

or specific symbols. Initial words, all capital words or first character capitalized (**f3**): in English, abbreviation words or words with all capital characters are probably the specific entities which can be people name, organization name, location name or nation name. Word with alphabets and numbers mixed (**f4**): It is observed some organization tends to have a name with alphabets and numbers mixed for easily memorized. The famous American company "3M" is an instance of such word. Punctuation (**f5**): Rather than engaging in deficiency of semantic unit segmentation, the punctuations in sentences are the best boundary of semantic unit that separate different phrases or clauses. Quotation mark is a punctuation that indicates the sentence of speech or subjective expressions.

B. Predicate-Argument Structure features

Predicate-argument structure (PAS) has been successfully implemented in semantic role labeling. PAS is a structure that captures the events of interest and the participant entities involved in events that correspond to predicate and arguments respectively. Generally speaking, predicate is usually a verb that conveys the type of event. The types and numbers of arguments are totally different while each predicate is different in syntactical or pragmatic view (e.g. transitive verb vs. intransitive verb).

Position of predicate (**f6**): arguments usually get closed to the predicate, especially agent and patient (subject and object of verb). The word "said" in the beginning or ending part of sentence possibly indicates the agent of speech event is located at beginning or ending part of the sentence.

Before or after predicate (relative position from predicate, **f7**): the arguments before or after predicate are totally different types of semantic roles. For example, in sentence "Peter chases John." Since the predicate is "chases" and two arguments Peter (arg0) and John (arg1) correspond to agent (Subject) and patient (object) respectively, it can be concluded the relative position of arguments from predicate impacts the semantic roles while syntactic categories are the same

Voice (**f8**): whether the predicate is active or passive voice that can affect the arguments type. In sentence "John is chased by Peter" the predicate changes the voice, not only the tense of verb is modified but both sequences of arguments are changed. If considering both relative position from predicate and voice, the resolution between opinion holders and the other labels is more feasible.

C. Syntactic features

Sub-categorization of predicate (**f9**): the feature is the verb phrase sub-structure that expressed the VP sub-parse tree structure where predicate located. Dash frame in Figure 1 indicates the sub-categorization of "decided" is "VP➔VBD-S". The feature helps analysis the phrase or the clause that follows predicate, and increases the ability of discriminating between arguments.

Head word and its POS (**f10**): the features are syntactic head of the phrase and the syntactic category of head. Different heads in noun phrase can be used to express different semantic roles. If the head word is "he" or "Bill" rather than "computer", then the probability that the noun phrase is the opinion holder increases. The Collins' head word algorithm is adopted for head word recognition [4]. Since the head of prepositional phrase (PP) is preposition and the significant semantic meaning in PP might be the noun phrase (NP), we also added the head of NP in PP as a feature which is called content word of PP. We also included the head word and head word's POS of parent node and grandparent node for considering the contextual syntactic features in linear data.

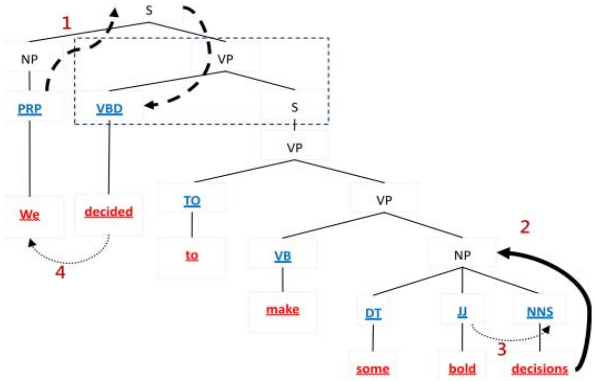

Figure 1 The syntactic features in a parse tree

Syntactic category of phrase (phrase type, **f11**): different semantic roles tend to be realized by different syntactic categories. The opinion holders are usually the noun phrases and sometimes prepositional phrases, but labels "*objective speech event*" and "*direct-subjective*" tend to be verb phrases. Our opinion holder identification model is realized by the linear CRF model, but the syntactic categories of the word or the phrase are not pure linear data which has been expressed in tree structure, i.e., the phrase type of word "decisions" in Figure 1 is NNS while the phrase type of "some bold decisions" is NP (i.e., the bold arrow). Instead of considering all syntactic categories in a syntactic parse tree which increases the space complexity dramatically, we traced upward only three syntactic categories of non-terminal nodes from the parent of leaf nodes if the head word of parent phrase is such leaf node. For example in Figure 1, because the head word of NP "some bold decisions" is "decisions" and the head of VP "make some bold decisions" is "make", the noun "decisions" would have phrase type only NNS and NP without VP. In contrast, the verb "make" would contains VB, VP and VP in sequence (The detail data view is illustrated in Figure 2 of Appendix). Hence, we embedded the syntactic features of phrase in head word of phrase to solve linear data structure problem. Considering contextual syntactic features by tracing bottom-up from the leaf of parse tree structure is better than by using sliding window method that merely adds POS of the previous and next token linearly without capturing global view of syntactic tree structure.

Syntactic path and partial path (**f12**): according to Gidea's statistical results [7], the path VB↑VP↓PP has 14.2% relative frequency to be PP argument or adjunct; path VB↑VP↑S↓NP↓ is 11.8% to be subject; path VBD↑VP↑NP↓ has 10.1% chance to be object of the sentence; and the VB↑VP↓ADVP is adverbial adjunct with 4.1%. The above path help predict the semantic labels. The syntactic path feature describes the syntactic relation from constituent to the predicate in the sentence with the syntactic categories of node passed through. In Figure 1, the path (i.e., the bold dash arrow) from "We" to "decides" can be represented as either "PRP↑NP↑S↓VP↓VBD" or "NP↑S↓VP↓VBD" depends on whether the constituent is PRP or NP of word "We". The deep parse tree can make the string of path too long, and would the data sparseness problem would happen. The partial path is part of syntactic path which contains the lowest common ancestor of constituent and predicate. (e.g., the lowest common ancestor is S in the sentence, so the partial path is reduced to "PRP↑NP↑S") Using partial path feature can solve the sparseness problem.

Based chunk (**f13**): the based chunk feature is similar to the phrase type feature but without the phrase type overlapped. In Table 2, the sentence S is consist of NP (We) and VP (decided to…), but this VP can be divided into non-overlapping sub-

Table 2 Sample based chunk features (f13)

We	*decided*	*to*	*make*	*some*	*bold*	*decisions*
B-NP	**B-VP**	**I-VP**	**I-VP**	**B-NP**	**I-NP**	**I-NP**

phrases which are combined by VP (decided to make) and NP (some bold decisions). We represent the based chunk in IOB format which makes the segmentation of phrase boundary more precise.

Subordinate noun clause followed verb and noun phrase before verb phrase (**f14**): Since our phrase type feature is only three levels of syntactic category from the parents of parse tree leaf nodes, the whole sentence structure information may be omitted if the parse tree is constructed deeply. In sentence "The management believed that …," the subordinate noun clause followed verb "believed" is usually embedded with subjective expressions and opinion targets. We used Stanford tregex toolkit to extract such pattern from the parse tree [21].

Syntactic dependency (**f15**): the purpose of dependency feature is not only to encode the syntactic structural information but also to capture the grammatical relation that includes three type of grammar dependency "subject relationship", "modifying relationship" and "direct-object relationship". Subject relationship includes "nominal subject" and "passive nominal subject", which correspond to the noun that is the syntactic subject of active and passive clause; when the governor of this relation is linking verb, the complement of the linking verb can be a noun or an adjective. Modifying relationship consists of adjectival modifier or adverbial modifier, which is any adjectival (adverb) word that modifies the meaning of noun (verb or adjective). Direct-object relationship indicates the noun that is the direct object of verb. We utilized Stanford dependency parser for the dependency features that indicate both govern word and dependent word in dependency relationship [21]. The opinion holders, opinion words in subjective expressions and opinion targets are correlated with the subject, modifying and direct-object relationship individually. We can get the whole picture from example sentence in Figure 1: the label of phrase "*to make some bold decisions*" is "*expressive-subjectivity*", we can observe that the opinion word in the phrase is "*bold*" with a adjective POS that modifies the noun "*decisions*". Since the word "*we*" is the subject of verb "*decides*", it suggests the subject relationship of verb can predict the opinion holders.

D. Simple semantic features

We employed dictionary based and statistical learning method to capture the simple semantic features.

Named entity recognition (NER, **f16**): utilizing syntactic features solely is hard to distinguish between the entity name and the others given a noun phrase. Although NER is not a comprehensive semantic feature, it can better separate person name which is possible the opinion holder or the opinion target from the others. Stanford NER [21] is employed to label people name, organization name and location name from other words. Further, the rare names which result in sparse feature data can be released.

Subjective word and its polarity (**f17**): the subjective words appear in sentences can help not only judging whether it is a opinion sentence but detecting the opinion words in the labels "*expressive-subjectivity*" and "*direct-subjective*". The subjective words can be classified into two aspects which are strength of subjective and polarity. According to different levels of subjective, the strength can be either one of objective, weak subjective and strong subjective. Moreover, the weak and strong subjectivity can still be divided into three kinds of polarity which are positive, negative and neutral. That is to say, there are 7 possible combinations of feature values. The subjective word dictionary used in the paper was manually collected by Wiebe [24].

Verb-clusters of predicate (**f18**): the similar semantic meaning verbs might appear together in the same document. For arranging the semantic related verb in a group, we use verb clusters to avoid the occurrence of rare verb which would deteriorate the model performance. ASSERT toolkit adopted probabilistic co-occurrence model to cluster the co-occurrence of verb into 64 clusters.

The frame of predicate in the FrameNet (**f19**): the FrameNet [6] is a corpus of sentences that has been hand-annotated for predicate and their arguments while predicates in the corpus are grouped into semantic frame around a target verb which has a set of semantic roles. Since every verb indicate different event of interest, the predicate-argument structure would be totally different. We used the FrameNet for querying of frame name that predicate belongs to.

IV. EXPERIMENTAIL EVALUATION

This section includes the experiment and evaluation of opinion holder identification CRF model. Furthermore, we also evaluated the model for different annotation labels mentioned in Section II.

A. Design of the experiments

The design of experiments is in Figure 3. Firstly, we preprocessed the MPQA corpus for making sure they can be parsed by the syntactic parser and are consist with IOB format. Secondly, we used linguistic features discussed in Section III to extract the feature values. Thirdly, we chose feature sets by different linguistic characteristics and transformed them in CRF data format. Finally, we trained and evaluated the CRF with different data sets and parameters.

The CRF model is implemented by sequential CRF tagging model of MALLET toolkit [15]. We trained and tested the CRF model with feature set which is discussed in Section II by 10325 sentences of the MPQA corpus, and the evaluation was performed on 30% holdout test data. The training iteration is 500 and the Gaussian variance is 10 with first-order CRF used.

The model is evaluated with respect to precision (***p***), recall (***r***) and $\boldsymbol{F_1}$ measure. The correct prediction of opinion labeling is defined as exactly matching between the CRF predicted label and the MPQA annotated label of specific phrase sequentially, and label of each token in the phrase should be the same. Definition of precision (***p***), recall (***r***), $\boldsymbol{F_1}$ measure and token accuracy (***a***) are described as,

***p* = the proportion of opinion labels predicted by the model is correct.**

***r* = the proportion of correct opinion labels is predicted by the model.**

$$\boldsymbol{F_1} = \frac{1}{\alpha\left(\frac{1}{p}\right)+(1-\alpha)\left(\frac{1}{r}\right)} \quad (1)$$

While α is the weight that controls the contribution percentage of precision and recall. That is to say, the $\boldsymbol{F_1}$ measure is a performance metrics that is the weighted harmonic mean of precision and recall. We set α to 0.5 for equally weighting the precision and the recall. Another loose performance metrics is token accuracy (***a***). Rather than comparing with exactly

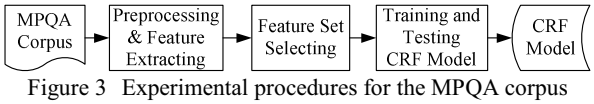

Figure 3 Experimental procedures for the MPQA corpus

Table 3 Results of different feature sets ("agent")

	Feature set	*a*	*p*	*r*	F_1
	Panel A no predicate-argument structure feature set				
A	f1+f2(**lemma**)	67.64	56.6	29.48	38.77
B	f1+f3~f5(**orthographical**)	63.76	53.49	22.49	31.67
C	f1+POS	64.03	66.42	16.85	26.88
D	f1+POS+f16(**NER**)	71.92	58.66	39.69	47.35
E	f1+POS+f15(**dependency, dep.**)	71.22	62.71	42.36	50.57
F	f1+f13(**based chunk**)	71.16	57.05	41.83	48.27
G	f1+f2+POS+f13+f15~f17	66.01	66.67	25.35	36.74
H	**f1~f5**+POS+f13+f15~f17	65.77	69.77	19.22	30.14
I	f1+f2+POS+f13+**f14~f17**	65.79	69.09	27.92	39.77
J	f1+f11(**phrase type**)	70.89	69.36	17.32	27.72
K	f1+f10(**head**)	70.67	27.07	4.27	7.37
L	f1+f11+f12(**phrase type and head**)	70.64	65.09	16.99	26.94
	Panel B expanded sentences by multiple predicates				
M	f1+f12(**only path**)	71.14	62.52	15.07	24.29
N	f1+f10+f12(**path and head**)	71.04	60.05	16.59	26
O	f1+f10+f11(**phrase type and path**)	71.02	68.92	21.14	32.36
P	f1,f2,f6~f19(**dep. excluded**)	70.71	64.01	50.6	56.52
Q	f1,f2,f6~f19(**NER excluded**)	70.88	69.91	35.91	47.45
R	f1,f2,f6~f19(**path excluded**)	71.23	68.62	35.73	46.99
S	f1,f2,f6~f19(**head excluded**)	71.04	64.69	48.25	55.27
T	f1,f2,f10~f17(**predicate excluded**)	71.04	69.4	38.02	49.12
U	f1,f2,f6~f19(**path, dep. excluded**)	70.64	67.9	32.15	43.63
V	f1,f2,f6~f19(**full**)	70.93	69.96	36.45	47.93
W	f1,f2,f6~f19(**no "target" label**)	76.97	70.84	38.28	49.7

matching criteria, token accuracy compares whether the predicted label and annotated label of single token are the same within token-level without considering the other tokens.

B. Experiment results

The feature set selection is applied for training and testing the CRF models (Table 3). Since there may be more than one predicate in a long sentence and some feature values related to predicate would be changed, we expanded one instance (i.e., one sentence data record) of the data (Panel A) into many instances with different feature values when there were more than one predicate in one sentence (Panel B).

We can observe from the feature set B (only orthographical feature) has the lowest token accuracy, and the feature set H (orthographical feature added) has significantly decreased the recall by about 6% when comparing with feature set G. On the other hand, what orthographical feature can capture is all served by NER features and POS of token. Hence, all the other feature sets are excluded the orthographical features excepted of set B and H.

The difference between feature set C (POS) and J (phrase type) is in granularity, it means the POS is merely the token-level syntactic category but the phrase type is the constituent-level syntactic category which is embedded in head word of the phrase. Since the linguistic characteristics of both are the same, the slight difference between their $\boldsymbol{F_1}$ measures is not out of expectation. Although their recalls are unfavorable among feature sets, the relatively higher precision is the reason to be reserved in following trials.

The feature set K (head feature) has the lowest $\boldsymbol{F_1}$ measure in all single feature sets which are tokens (f1) plus single feature in Panel A, and accordingly, feature set S which has excluded the head feature enhanced the recall promisingly but also accompanied by the drop in the precision. We inferred that the head is not a good indicator for phrase boundary detection but would make the model be conservative in predicting which constrains the recall when head feature included, so releasing head feature (set S) makes the recall soar up.

The combination of token, POS and dependency features (E) has the best $\boldsymbol{F_1}$ measures in Panel A, but there is an anomaly that dependency features excluding from full feature sets in Panel B also results in promising performance with highest recall. Because feature set U without both path and dependency would cause the lower performance.

Table 4 Performance with explicit expression sentences

Annotation labels	**all sentences**			**explicit expression**		
feature set	**G: (f1+f2+POS+f13+f15~f17 with "target" label)**					
%	***p***	***r***	$\boldsymbol{F_1}$	***p***	***r***	$\boldsymbol{F_1}$
Agent	66.67	25.35	36.74	64.34	43.61	51.98
Obj.speech-event	36.89	9.87	15.57	44.12	36.36	39.87
Direct-subjective	44.61	14.75	22.17	51.09	32.99	40.09
Expressive-sub.	7.65	0.8	1.46	27.67	10.49	15.21
target	0.65	0.1	0.17	10.19	5.37	7.03
Other	37.43	51.23	43.25	48.29	57.42	52.46
Average	37.72	44.48	40.82	48.09	52.2	50.06
Token accuracy	66.01			68.79		

Table 5 Performances achieved by the 1st and 2nd order CRF

Annotation labels	**First-order CRF**			**Second-order CRF**		
feature set	**W (explicit expression and without "target")**					
%	***p***	***r***	$\boldsymbol{F_1}$	***p***	***r***	$\boldsymbol{F_1}$
Agent	70.54	41.95	52.61	69.29	39.17	50.05
Obj. speech-event	40	5.88	10.26	46.15	5.43	9.72
Direct-subjective	44.22	18.38	25.97	45.07	17.2	24.9
Expressive-sub.	18.06	3.43	5.76	8.15	1.45	2.46
Other	32.97	42.12	36.99	30.68	39.62	34.58
Average	34.12	39.09	36.43	31.78	36.68	34.05
Token accuracy	73.52			72.31		

In summary, we can conclude that feature sets with lemma, orthography, POS, phrase type and head are relatively inferior to feature sets with NER, syntactic dependency and based chunk by observing the $\boldsymbol{F_1}$ measures among feature set A, B, C, F, J and K when features consist of f1 plus another feature. By comparing the performance among feature set G, H and I, the inclusion of feature f14 is better than inclusion of orthographical features under no predicate-argument structure situation. The $\boldsymbol{F_1}$ measure of full feature set (V) is not significantly different from sets R, Q and T, but is explicitly higher than P, S and U; we can draw the conclusion that the exclusion of feature in NER and path from full set would not affect the system performance but exclusion of feature in head would.

In Table 4, we set the "explicit expression" condition to select sentences from all MPQA sentences while explicit expression is defined as: both labels "*agent*" and one of "*expressive-subjectivity*", "*objective speech event*" and "*direct-subjective*" occur in the same sentence. The explicit expression sentences include 4823 sentences and 1447 sentences held for testing. Since the sentences are all explicit expression sentences, the performance in right columns all outperform the left columns. Accordingly, the predictability of syntactic structure and predicate-argument structure in explicit expression sentences is better than in implicit expression sentences. In Table 5, we compared the first-order CRF performance with second-order CRF performance under explicit expression sentences, but the second-order performance is not been improved. The probable reason is that the second-order model is over-fitted to training data whose size is too small to generate a representative model.

In the left side column of Table 6, the performance of annotation label "target" is worst in all annotation labels. The reason of such result is that the frequency of "target" occurrence in all MPQA samples is small and the length of target is longer than the other annotation labels, which are inherent problems caused when annotated. Removing the noise label "target" facilitate the performance of other annotation labels except for the precision of label "*Objective speech-event*". Average speaking, the recall without targets is outperformed the recall with targets.

Table 6 Performance with and without "target" label

Annotation labels	With target label			Without target label		
%	*p*	*r*	F_1	*p*	*r*	F_1
Agent	63.68	27.99	38.89	72.29	30.81	43.21
Obj.speech-event	42.86	1.07	2.09	29.07	4.5	7.8
Direct-subjective	46.69	9.11	15.25	56.23	11.96	19.72
Expressive-sub.	14.18	1.74	3.1	27.67	5.01	8.48
target	4.88	0.19	0. 36	-	-	-
Other	40.86	52.45	45.94	46.98	56.37	51.25
Average	41.03	47.81	44.16	47.19	52.4	49.66
Token accuracy	75.08			81.27		

V. MINING OPINION PATTERNS FROM FINANCIAL STATEMENTS

The main purpose of this section is automatically extracting the opinion holders and subjective expressions in financial statements by CRF model trained in previous section without manually filtering. For both identifying and extracting the opinion patterns in the U.S. financial statements, the text in financial statements should be transferred into the format which our system can handle with. Besides, the focus of our work is the textual contents of 10-K, so the process of eliminating redundant contents and useless supplements is necessary.

We preprocessed the filings for extracting footnote section and got rid of redundant information. The following is cleaning preprocesses: first of all, we removed the HTML tags, pictures, tables, front and ending matters and exhibitions for keeping the useful item sections; the next, we dropped the line contains too much white spaces, symbols, numbers or non-meaning words (e.g., fragment left after elimination) by regular expression; finally, we adopted LingPipe sentence model [12] to segment the filings into sentences for applying the opinion holder identification model which is sentence-based method, and the total sentence number after preprocessing is 1.3 million. Instead of using 1.3 million sentences which are not all opinion sentences, we selected 85,394 sentences only if the sentence contains both the financial positive/negative words and words list in MPQA "*objective speech event*" and "*direct-subjective*" (e.g., "believe"). The financial positive/negative words which were collected by Loughran and McDonald [14], and all of them are unigrams (e.g., single word "failure"). Finally, we chose 30,381 sentences that can be processed by syntactic parser, and restricted the sentence length ranged from 8 to 100 words.

The goal of the automatic extraction is to explore the opinion holders which are the entities that involved in financial statements and also the key-phrases in subjective expressions that opinion holders have expressed. We chose feature set W (Table 3) which has the highest precision in unseen data. For better extraction and centralized statistic figures, we replaced every token in sentences with their lemma word and then substituted person name, organization name and location name with PERSON, ORGANIZATION (ORG.) and LOCATION (LOC.) correspondingly.

In Panel A of Table 7, there are top 8 frequent participant entities and their subjective expressions in above sample sentences. The financial statement is the financial report of the specific corporation, so words "we" and "the company" and "management" reference to the same entity and they must be the most frequent words while "the company" are the subject of the report and also "we" and "management" are the responsible entity who compile the report. We can find "the" in agent list is the segmentation error which wrongly split the noun phrase at definite article.

The main contribution of the work is to automatically extract the opinion patterns in form of multi-word expressions (MWEs) which are important in financial decision making, such as the subjective expression phrase "scientifically feasible commercially viable opportunity" which suggests the positive investment outlook and the phrase "substantially doubt about its ability to continue as a going concern" which indicates the bankruptcy in the near future (in Panel B of Table 7). Our CRF model can successfully extract the useful subjective MWEs, while the unigram (single-word) model cannot capture as much semantic information as what our models can.

Table 7 Opinion patterns extracted from 10-K filings

Panel A Top 8 frequent phrases					
Agent list	**Freq.**	**direct subjective**	**Freq.**	**Subjective expression list**	**Freq.**
We	8249	believe	5352	may not be able	140
the company	1840	agree	1274	may not be recoverable	137
the ORG.	948	expect	888	reasonably assure	62
management	606	**cannot assure you**	635	may be impaired	55
the	408	intend	546	substantial doubt	51
It	394	do not believe	538	would not be able	49
company	328	provide	466	may not be successful	48
PERSON	249	determine	346	would become exercisable	45
Panel B Some of the other frequent phrases					
the plaintiff	76	anticipate	169	scientifically feasible commercially viable opportunity	22
the executive	71	deny	143	could adversely affected	19
the debtor	35	conclude	130	could significantly reduce our revenue	16
the credit agreement	31	violate	19	substantially doubt about its ability to continue as a going concern	15

VI. DISCUSSIONS AND CONCLUSIONS

We present an application of linear-chain CRF models which embrace features in morphology, orthography, predicate-argument structure, syntax and simple semantics for opinion holder identification and opinion patterns extraction. The CRF models for opinion holder identification achieved their best performance, i.e., 0.72 for precision and 81% for token-based accuracy, when we ran experiments with the MPQA corpus. We also extracted opinion holders and subjective expressions from U.S. financial statements with our CRF models, and were able to find interesting patterns algorithmically. The major contribution of our work is finding ways to automatically expand the lists of words that are linked to positive and negative financial statuses into subjective multi-word expressions (MWEs) successfully.

There are some challenges in automatic opinion holder identification. First of all, an individual sentence can contain more than one opinion holders, and different opinion holders may express different expressions which might cause the nested structure problem. Since the opinion agents "*Datanalisis' November poll*" and "*55.3 percent*" in Table 1 are nested, the opinion agents and expression matching is a hard nut to crack. Secondly, anaphor resolution is still hard to solve. To avoid the same words appearing repeatedly in statements, pronouns and abbreviations are frequently used. It is our goal to find the identities referred by these substitute words. Finally, co-reference problem is another barrier. How could our algorithms know that both "Intel corp." and "the leading semiconductor company" refer to the same company? The problems described above are beyond the prediction competence of our CRF models that capture mainly the syntactic features of sentence-level and word-level features but no comprehensive semantic features.

Some constraints arise while attempting to solve non-linear

problems with linear machine learning methods and conditionally independent assumptions. First of all, the dependency between features may cause the estimation of parameters in CRF models biased. Moreover, our linear CRF models cannot capture the whole syntactic structure information that is expressed in parse tree structures. Finally, the conditionally independent assumption blocks the influences of the long distance dependency phenomenon. In the future, the principle component analysis (PCA) would be taken to tackle the feature dependency between syntactic features, and the arbitrary structure of CRF may relieve the long dependency problem. Identifying the opinion holders and extracting opinion patterns from the sentences are the main tasks of the paper. The opinion patterns in financial statements can be used to verify whether the opinion patterns are indicative of the future financial performance of the corporations. It is also interesting to examine whether the sentiment of the opinion patterns agrees with the financial ratios reported in the current and future financial statements.

ACKNOWLEDGMENT

The research has been partially supported by research contracts NSC-99-2221-E-004-007 and NSC-97-2221-E-004-007-MY2 of the National Science Council of Taiwan.

REFERENCES

[1] W. Antweiler and M. Z. Frank, "Is all that Talk just Noise? The Information Content of Internet Stock Message Boards," *J. of Finance,* 59(3), 1259-1294, 2004.

[2] Automatic Statistical SEmantic Role Tagger-v0.14b (ASSERT), http://cemantix.org/assert.html.

[3] Y. Choi, C. Cardie, E. Riloff and S. Patwardhan, "Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns," *Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing,* 355-362, 2005.

[4] M. J. Collins, *Head-Driven Statistical Models for Natural Language Parsing*, Ph.D. thesis, Univ. of Pennsylvania, 1999.

[5] Electronic Data Gathering, Analysis and Retrieval system (EDGAR), http://www.sec.gov/edgar.shtml.

[6] FrameNet, http://framenet.icsi.berkeley.edu/.

[7] D. Gildea and D. Jurafsky, "Automatic Labeling of Semantic Role," *Computational Linguistics*, 28(3), 245-288, 2002.

[8] Illinois Chunker, http://cogcomp.cs.illinois.edu/page/software.

[9] S.-M. Kim and E. Hovy, "Identifying Opinion Holders for Question Answering in Opinion Texts," *Proc. of AAAI Workshop on Question Answering in Restricted Domains*, 20-26, 2005.

[10] J. D. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proc. of the Int'l Conf. on Machine Learning,* 282-289, 2001.

[11] F. Li, "Do Stock Market Investors Understand The Risk Sentiment Of Corporate Annual Reports?" Univ. of Michigan Working Paper, 2006.

[12] LingPipe 3.9 sentence Model, http://alias-i.com/lingpipe.

[13] B. Liu, "Sentiment Analysis and Subjectivity," *Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau (editors), CRC press , Second Edition, 2010.

[14] T. Loughran and B. McDonald, "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *J. of Finance*, 66(1), 67-97, 2011.

[15] MAchine Learning for LanguagE Toolkit-2.0.6 (MALLET), http://mallet.cs.umass.edu.

[16] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval,* Cambridge Univ. Press, 2009.

[17] Multi-Perspective Question Answering 2.0 (MPQA), http://www.cs.pitt.edu/mpqa.

[18] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 79-86, 2002.

[19] S. Pradhan, W. Ward, K. Hacioglu, J. Martin and D. Jurafsky, "Shallow Semantic Parsing Using Support Vector Machines," *Proc. of the Human Language Technology Conf./North American Chapter of the Association of Computational Linguistics*, 2004.

[20] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 25-32, 2003.

[21] Stanford NLP Toolkits, http://nlp.stanford.edu/software.

[22] P. C. Tetlock, M. Saar-Tsechansky and S. Macskassy, "More than Words: Quantifying Language to Measure Firms' Fundamentals," *J. of Finance,* 63(3)*,* 1437-1467, 2008.

[23] J. Wiebe, R. Bruce and T. O'Hara, "Development and Use of a Gold Standard Data Set for Subjectivity Classifications," *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 246-253, 1999.

[24] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing*, 347-354, 2005.

Appendix

Feature-Level		Set #	Feature-values (partially selected feature set)						
Token-level features	Features of leaf node	f1	We	decided	to	make	some	bold	decisions
		f2	we	decide	to	make	some	bold	decision
		f17	objective	weak, neutral	objective	objective	objective	strong, positive	objective
		f7	before		after	after	after	after	after
Phrase-level features	Syntactic features of leaf node	f10	PRP: we		TO: to	VB: make	DT: some	JJ: bold	NNS: decisions
		f11	PRP		TO	VB	DT	JJ	NNS
		f12	PRP↑NP↑S		TO↑VP↑S↑VP	VB↑VP↑VP↑S↑VP	DT↑NP↑VP↑VP↑S↑VP	JJ↑NP↑VP↑VP↑S↑VP	NNS↑NP↑VP↑VP↑S↑VP
	Syntactic features of leaf node's parent	f10	PRP: we	PREDICATE	O	VB: make	O	O	NNS: decisions
		f11	NP		O	VP	O	O	NP
		f12	NP↑S		O	VP↑VP↑S↑VP	O	O	NP↑VP↑VP↑S↑VP
Sentence-level features		f8	active		active	active	active	active	active
		f9	VP➔VBD-S		VP➔VBD-S	VP➔VBD-S	VP➔VBD-S	VP➔VBD-S	VP➔VBD-S
IOB format			B-	B-	B-	I-			
LABELS			agent	direct-subjective	expressive-subjectivity				

Figure 2 Example sentence expressed in linear CRF data view