# ACRank: a multi-evidence text-mining model for alliance discovery from news articles

Yilu Zhou
*Fordham University, New York, New York, USA, and*
Yuan Xue
*Pennsylvania State University, University Park, Pennsylvania, USA*

Received 6 June 2018
Revised 23 November 2019
17 February 2020
3 May 2020
Accepted 13 May 2020

## Abstract

**Purpose** – Strategic alliances among organizations are some of the central drivers of innovation and economic growth. However, the discovery of alliances has relied on pure manual search and has limited scope. This paper proposes a text-mining framework, ACRank, that automatically extracts alliances from news articles. ACRank aims to provide human analysts with a higher coverage of strategic alliances compared to existing databases, yet maintain a reasonable extraction precision. It has the potential to discover alliances involving less well-known companies, a situation often neglected by commercial databases.

**Design/methodology/approach** – The proposed framework is a systematic process of alliance extraction and validation using natural language processing techniques and alliance domain knowledge. The process integrates news article search, entity extraction, and syntactic and semantic linguistic parsing techniques. In particular, Alliance Discovery Template (ADT) identifies a number of linguistic templates expanded from expert domain knowledge and extract potential alliances at sentence-level. Alliance Confidence Ranking (ACRank)further validates each unique alliance based on multiple features at document-level. The framework is designed to deal with extremely skewed, noisy data from news articles.

**Findings** – In evaluating the performance of ACRank on a gold standard data set of IBM alliances (2006–2008) showed that: Sentence-level ADT-based extraction achieved 78.1% recall and 44.7% precision and eliminated over 99% of the noise in news articles. ACRank further improved precision to 97% with the top20% of extracted alliance instances. Further comparison with Thomson Reuters SDC database showed that SDC covered less than 20% of total alliances, while ACRank covered 67%. When applying ACRank to Dow 30 company news articles, ACRank is estimated to achieve a recall between 0.48 and 0.95, and only 15% of the alliances appeared in SDC.

**Originality/value** – The research framework proposed in this paper indicates a promising direction of building a comprehensive alliance database using automatic approaches. It adds value to academic studies and business analyses that require in-depth knowledge of strategic alliances. It also encourages other innovative studies that use text mining and data analytics to study business relations.

**Keywords** Strategic alliances, Knowledge discovery, Business intelligence, Web mining, Text mining, Information extraction, Template-based, Chunk parsing

**Paper type** Research paper

## 1. Introduction

Inter-firm collaboration has exploded in the past few decades. The nature of collaboration has shifted from peripheral interests to the very core functions of cooperation and from equity to non-equity forms. This phenomenon has drawn strong interest from analysts, business strategists, and policymakers from various fields, including economics, management, public administration, science, and technology (Vonortas *et al.*, 2003).

Strategic alliances are defined as "contractual asset pooling or resource exchange agreements between firms" (Stuart, 1998). It is a widely studied topic in strategic management. A solid strategic alliance database would allow policy, innovation, and

This material is based upon work supported by the National Science Foundation under Grant No. CNS1551004.

Information Technology & People
Vol. 33 No. 5, 2020
pp. 1357-1380
© Emerald Publishing Limited
0959-3845
DOI 10.1108/ITP-06-2018-0272

economics researchers to better understand this growing phenomenon. Many existing databases, however, are narrowly focused and cannot link with other databases; many of the private-sector databases are also too costly for researchers and students to access; and traditional data-collection techniques and primary sources have been limited. The exclusive reliance on a limited set of popular sources for related information also creates bias. Candidly, these shortcomings also apply to the most frequently used and cited data based on research partnerships (Hagedoorn *et al.*, 2000). Policymakers, academics, and businesses have to deal with the incomplete, outdated, and expensive products currently available.

The main challenges in constructing a better knowledge repository are not different from many other fields: information overload and the limitations of human cognition. Currently, the discovery of strategic alliances relies on humans physically reading news articles and company reports and then inputting data manually, strictly limiting shared knowledge, including the number of alliances that can be searched for (thus decreasing completeness), the speed of knowledge updates, and increased cost (paying for researchers to manually comb through records).

An emerging trend of business intelligence, a form of information technology, is using the Web as arepository to study strategic relations between organizations, such as competitions and alliances. It faces several challenges, however. First, alliance announcements are not aggregated centrally, appearing in many places, such as news articles, trade journals, and government filings. Second, the appearance of a valid alliance is rare compared to the number of documents that require scanning. Third, alliance partnerships can take a variety of forms, such as joint ventures, research cooperation, service cooperation, and informal agreements, increasing the difficulty of seeking them manually using predefined keywords. Last but not least, the framework needs to effectively handle large amounts of unstructured textual data and accurately identify alliances. However, in these fields, few holistic frameworks are available to deal with a real-world knowledge extraction problem such as strategic alliance extraction.

We aim to address the limitations of manual work by developing an intelligent knowledge-extraction framework to extract existing alliances from open resources, such as published news articles. We aim to design and develop and IT artifact what could (1) support human analyst by offering a wider coverage of strategic alliances with reasonable precision, (2) allow more efficient alliance identification with the help of the framework. Following the Design Science paradigm, the framework comprises unique components tailored to alliance extraction: meta-search, dependency parsing, entity extraction, relation extraction, and information integration. We evaluated the effectiveness of our framework in a case study and compared the coverage of our approach with the gold standards: the Thomson Reuters SDC database and expert judgment. Our research is a first step toward building an alliance knowledge repository and thus provides rich evidence for strategic alliance–formation studies.

We first review the characteristics of strategic alliance, current approaches to support strategic alliance discovery, and the field of text mining. We then describe the challenges associated with finding strategic alliances from massive amounts of documents and present our alliance knowledge-extraction framework. Following our framework, we evaluate the system through a case study. Finally, we conclude the paper explaining its contributions and future directions.

## 2. Research background

With technological innovation and diffusion being the primary driver of today's knowledge economy and alliances becoming an ever-increasing source of technological innovation and diffusion, economists, managers, and policymakers need access to information on those alliances' successes and failures. This paper focuses on formal strategic alliances, not other types (e.g., technology-focused partnerships). Table 1 summarizes a widely-used classification scheme of strategic alliances based on the relations between participating companies (Nooteboom, 1999; Yoshino and Rangan, 1996). These numerous forms of alliances increase the difficulty of finding alliances manually.

### *2.1 Current methods to identify alliances*

Table 2 summarizes each approach to identifying alliances with their pros, cons, and data sources. Co-authorship analysis is a promising way to identify research and technology partnerships (Joly and de Looze, 1996; Tsuji, 2002), but patenting levels vary by sector and firm size, among other factors, so any alliance data based on that metric may be biased to sectors with strong patenting traditions or needs. Filings with governments can provide accurate results and have been used to study alliances across countries (Hall *et al.*, 2001; Oxley and Wada, 2009), but this method largely depends on the availability of filings and has low alliance coverage. Press and trade publication analysis often work sin small-scale case studies that identify the major partners of one or a few companies (Yan *et al.*, 2016), but tremendous noise and extensive human laborundermine its effectiveness without the use of automated methods. Surveying alliance researchers and industry experts is another popular approach (Heimeriks and Duysters, 2007), but they rarely cover all firms in a sector, and their accuracy is contingent upon getting the right person—or group of people—to answer questions at the firm level. Patent analysis and survey approaches are unsuited to situations when policymakers need to know how generic innovation policies affect all players, and survey or patent analysis approaches can bias economists' perceptions of the inputs and outputs associated with operating an economy efficiently.

### *2.2 Available alliance knowledge databases*

As suggested by Schilling (2009), none of the existing databases maintained manually are considered accurate reflections of the entire population of strategic alliances, only of subsets. The Thomson Reuters SDC Alliance database is the most-cited and is populated by information manually extracted from newspapers, trade journals, government filings, and other press and news wire services. Thomson Reuters employees read these sources and, upon locating an alliance announcement or information relating to an alliance already in the database, update the database as needed. This data-entry has a time lag of many years, however. Other alliance databases are available (Table 3). However, they provide even less coverage of alliances by only focusing on specific industries and alliance types.

| Alliance type | Participating firms | Example |
|---|---|---|
| Horizontal strategic alliances | Competitors | Microsoft and IBMpartnership |
| Vertical strategic alliances | Upstream/downstream partners | Toyota's partnership with its suppliers |
| Intersectional alliances | Companies with little similarity/ connections | Barnes and Nobleand Starbucks partnership |
| Joint ventures | Companies forming a new company | Google and NASA collaboration on Google Earth |
| Equity alliances | Companies with shareholding relations | Panasonic and Tesla supply agreement |
| Non-equity strategic alliances | Companies with contractual relations | Starbucks and Kroger partnership |

**Table 1.**
Strategic alliance types

| Approach | Summary | Sources |
|---|---|---|
| Patent data, bibliometrics | *Pros:*<br>High data accessibility<br>Clean, measurable data<br>*Cons:*<br>Missing innovative activities<br>Low data reliability because of varied filling requirements<br>Missing alliances not reported in English<br>Missing alliances not reported in publications | USPTO; PubMed, Web of Science, etc. |
| Filings with government offices | *Pros:*<br>High data reliability<br>*Cons:*<br>Lacking comprehensive information<br>Lacking comparability between countries<br>Limited applicability to only specific alliances types | SEC; IRS; Dept. of Justice |
| Popular press, trade publications | *Pros:*<br>High data accessibility<br>*Cons:*<br>Limited focus on only documents in English<br>Labor-intensive manual work<br>Information overload | Collections of specific trade magazines and news publications |
| Surveys | *Pros:*<br>Quantifiable data<br>Specific/in-depth information<br>*Cons:*<br>Low response rate<br>Lacking generalizability | Administered to firms, universities, government funders, and laboratories |

**Table 2.**
Overview of research strategies for strategic alliances

| database | Size | Scope | Comments |
|---|---|---|---|
| Thomson Reuters SDC | Over 500 publications, corporate records, etc. | Any alliance in any sector might be reported; no lower limit on value of alliance | Considered the "gold standard" for existing alliance databases |
| MERIT-CATI (now UNU-MERIT) | English-language publications for non-US firms | Only records alliances with at least two industrial members; no government or university alliances | Does not cover US firms |
| CORE/NCRA | Department of Justice filings by private firms | Only those firms that register alliances with Dept. of Justice | No longer maintained; fewer firms filing alliances with Dept. of Justice |
| RECAP | Over 30,000 alliances and detailed corporate reports | Biotechnology alliances only | Lack of data in other sectors |
| BIOSCAN | Tracks 2,000 specific firms over time | Biotechnology alliances only | Excellent details on these specific firms; lack of data in other sectors |

**Table 3.**
Overview of existing strategic alliance databases

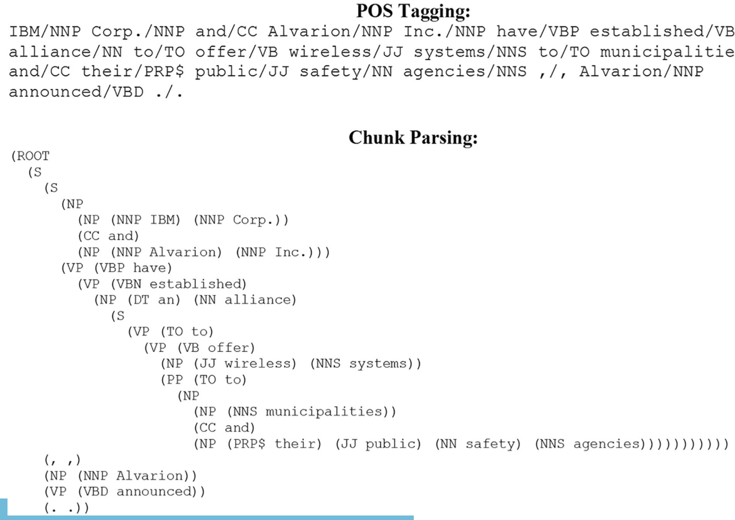
*2.2.1 Part-of-speech tagging.* POStagging is the process of assigning a part-of-speech tag to each word/token in a sentence. The POS tags consist of coded abbreviations conforming to the Penn Treebank scheme, the linguistic corpus developed by the University of Pennsylvania [1]. We also performed*chunk parsing* on the sentence text, which allowed grouping sentence tokens into larger chunks, each chunk corresponding to a syntactic unit such as a noun phrase or a verb phrase, helping both identify named entities and extract alliance-related verb phrases. Both POS tagging and chunk parsing can be performed using rule-based or statistical, learning-based approaches (Ratnaparkhi, 1996). Several open-source tools are available to perform English POS tagging and chunk parsing, such as OpenNLP [2] and LingPipe [3]. POS tagging and chunk parsing are considered syntactic analysis and shallow parsing in the natural language processing (NLP) field (Molina and Pla, 2002). We ran both types of parsing on the entire document set.

Figure 1 shows an example of POS tagging and chunk parsing results from analyzing the sentence "IBM Corp. and Alvarion Inc. have established an alliance to offer wireless systems to municipalities and their public safety agencies, Alvarion announced."

*2.3 Text mining and text mining–based applications in studying strategic relations*

*Text mining*, the process of discovering knowledge and trends from unstructured text (Tan, 1999), is a promising technique that could potentially automate alliance extraction, as it can handle large volumes of unstructured data (Fan *et al.*, 2006). *Information Extraction* (IE) is another important text-mining technique by screening out noise and extracting only structured information, such as entity names and their relations, from unstructured text, which can then be used for knowledge discovery (Nahm and Mooney, 2002; Mooney and Bunescu, 2005).

Text-mining performance depends on sentence-parsing, of which three levels exist: (1) Bag-of-Words (BoW) parsing, (2) syntactic parsing, and (3) semantic parsing. A simple BoW approach suffers from high noise and dimension. Syntactic parsing, also called "shallow parsing," includes techniques such as Part-of-Speech (POS) tagging, noun phrasing, and chunk parsing. Semantic parsing, also known as "deep parsing," is the most advanced method. It represents a sentence with a dependency parse.



**POS Tagging:**

```
IBM/NNP Corp./NNP and/CC Alvarion/NNP Inc./NNP have/VBP established/VBN an/DT
alliance/NN to/TO offer/VB wireless/JJ systems/NNS to/TO municipalities/NNS
and/CC their/PRP$ public/JJ safety/NN agencies/NNS ,/, Alvarion/NNP
announced/VBD ./.
```

**Chunk Parsing:**

```
(ROOT
  (S
    (S
      (NP
        (NP (NNP IBM) (NNP Corp.))
        (CC and)
        (NP (NNP Alvarion) (NNP Inc.)))
      (VP (VBP have)
        (VP (VBN established)
          (NP (DT an) (NN alliance)
            (S
              (VP (TO to)
                (VP (VB offer)
                  (NP (JJ wireless) (NNS systems))
                  (PP (TO to)
                    (NP
                      (NP (NNS municipalities))
                      (CC and)
                      (NP (PRP$ their) (JJ public) (NN safety) (NNS agencies))))))))))
    (, ,)
    (NP (NNP Alvarion))
    (VP (VBD announced))
    (. .))
```

**Figure 1.**
Example of POS tagging and chunk parsing

Domain-knowledge integration also plays an important role in text mining (Tan and Lai, 2000) by both reducing noise and keeping relevant information using input from domain experts (D'Haen *et al.*, 2016).

Most studies of text mining, however, focus only on identifying competitive relations from single news sources. Other studies of cooperation and alliance neither distinguish between alliances types nor rank extracted alliances by confidence. Alliance extraction is a more challenging problem than general relations-extraction between two companies because alliance evidence is sparse and takes a variety of forms. Thus, alliance-extraction techniques need integration with domain knowledge.

## 3. Research questions

Data limitations greatly affect alliance research. Collecting all alliance data would allow not only a more central repository of information but also research questions involving multiple areas of investigation by providing an overall picture of strategic alliances.

Despite the development of text mining and its wide applications in many scientific fields, to the best of our knowledge, no collaborations have been formed between the alliance and text-mining research communities. To bridge this gap and meet the challenge of building a text mining–based alliance extraction framework, we pose the following research questions:

(1) How can we design a framework to automate the discovery of strategic alliances from massive amounts of textual data?

To answer this question, we are specifically interested in looking into the following three aspects: 1) finding the data sources; 2) identify the necessary steps in the framework; 3) identify the appropriate text-mining techniques and theories to extract true strategic alliances from high volumes of noisy textual data; and 4) A ranking algorithm based on quality of alliances.

(2) Can our automated alliance discovery framework achieve adequate precision and recall compared to human analysts?

To answer the second question, we are interested in two types of benchmarks. For a direct comparison, if our approach is applied to the same dataset of news as what human analyzes, will our approach extract alliances with reasonable precision and recall? Second, compared to commercial databases, which was hand-crafted by many domain experts over time, will our approach be a good complimentary tool that address the coverage concern of current databases (Schilling, 2009).

## 4. An alliance knowledge-extraction framework

To bridge research gaps in strategic alliance discovery, we designed and built an automatic knowledge-extraction framework to identify alliances. Our initial study showed that alliance extraction from free text is even more challenging than other types of IE tasks because, for one, strategic alliance information can appear in any type of news but not often. Our experts found that 99% of Google results were noise. This lack of alliance information in news articles makes some well-performing learning-based algorithms, such as Support Vector Machine (SVM), fail. This problem indicated that the construction of a set of relevant documents with an abundant amount of strategic alliances was essential to extraction performance. The documents being parsed by our framework needed to be both comprehensive enough to cover most recently announced alliances and focused enough to avoid unnecessary noise. Conversely, staying at the syntactic level of document-processing would not meet our needs.

With these special challenges in mind, we designed the alliance extraction framework with the components shown in Figure 2. The final result of our text-mining model is a ranked list of possible alliances. We later discuss each component in detail.

### *4.1 Alliance news document collection*

The first step was collecting news documents containing company alliance information. The corpus we used for alliance identification included news documents mentioning strategic alliances formed by the company of interest in our analysis. We implemented a lexicon-based retrieval by first having domain experts create a lexicon containing keywords and phrases indicative of alliance relations, such as “partner” and “joint venture”. The terms in our lexicon were proposed and scrutinized based on the experience and knowledge of business experts, since overly general terms would have lowered the precision of the results and overly specific terms would have lowered the recall rate. Next, we implemented a meta-crawler to retrieve the relevant text documents indexed by major search engines, such as LexisNexis, Google News, and Thomson Reuters, using a combined search for both the names of companies of interest and lexicon keywords. When multiple search engines were used for meta-crawling, the results of different engines were merged into a single dataset with duplicate titles and contents removed. The purpose of using search engines that indexed news from multiple sources was to avoid the bias of relying on one news source and to obtain a more comprehensive coverage of possible strategic alliances (Chen *et al.*, 2002; Lawrence and Giles, 1999).

### *4.2 Pre-processing*

The resulting documents from the focused meta-crawling were sent to pre-processing, which involved the following three steps.

*4.2.1 Document indexing.* Document indexing involved data cleaning, document tokenization, and meta-data extraction. Meta-data included the source of publication, author, date of publication, length of the article, etc. These features are used in later steps of our analysis. Word and sentence positions were also recorded. For a uniform management of data from various sources, we convert all source files into XML format to create to access specific alliance evidence, such as specific dates or authors, information without which validating database accuracy is difficult.

*4.2.2 Entity extraction.* Entity extraction extracts and classifies rigid designators (Nadeau, 2007). Entity extraction can be done using a rule-based approach, a machine learning–based approach, or a hybrid approach. A rule-based approach focuses on extracting names using many rule sets designed by linguistic experts. A machine learning–based approach employs

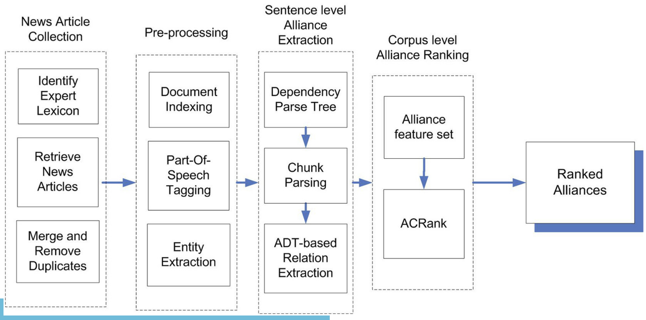


**Figure 2.**
Proposed framework for alliance discovery

a classification statistical model to solve the entity-recognition problem (Mansouri *et al.*, 2008). Our framework adopts a hybrid approach, combining rule-based and machine learning–based methods and thus benefitting from the advantages of each (Srihari, 2000).

Because of our unique domain, we first established an organization name list comprising publicly traded companies, major universities, and government organizations. We then developed an entity-extraction rule set capturing features of organization names such as capitalization and use of "Inc." and "Corp." Finally, we combined these results with a machine learning–based entity-extraction system such as those developed by OpenNLP [4] and the Stanford Natural Language Processing Group (Manning *et al.*, 2014). In OpenNLP, entity names are extracted using the Maxent library (Tsuruoka, 2006), which implements a maximum entropy model (Manning *et al.*, 1999). The Stanford CoreNLP tool recognizes entity names using a conditional random fields classifier (Lafferty *et al.*, 2001). Entity extraction performance was critical in our system, as it affected later alliance extraction quality. This multi-strategy approach adapts well to the alliance domain in our analysis by helping our system cover a wider range of organizations.

### *4.3 Sentence-level alliance extraction*

After document pre-processing, the next two steps focus on the extraction of alliance relations. In our alliance extraction framework, sentence-level alliance extraction tackles the problem from an IE point of view by asking only whether each document and sentence contains valuable alliance information or not. In corpus-level alliance ranking, it first turns unstructured text into structured features and then predicts the likelihood of each extracted relation being about a strategic alliance based on both document and sentence-level features.

*4.3.1 Dependency parse trees.* During sentence-level alliance extraction, each sentence that contains a potential alliance is an *alliance instance*. We used a *dependency parse tree*–based approach here, which presents different types of word dependencies organized in a hierarchical structure based on the similarities in their grammatical roles in sentences. While POS tagging and chunk parsing analyze syntactic structure, dependency parse trees analyze the parsing of the semantic structure, called "deep parsing" (Bunescu and Mooney, 2005; Culotta *et al.*, 2006; Frank *et al.*, 2003; Miller *et al.*, 2000; Zelenko *et al.*, 2003). To reduce time and resource costs, we reduced the number of target sentences by only parsing the sentences containing at least one organization name.

Figure 3 shows a dependency parse tree generated from the same sentence as Figure 1. While POS tagging captured the syntactic structure of the sentences, the dependency parse

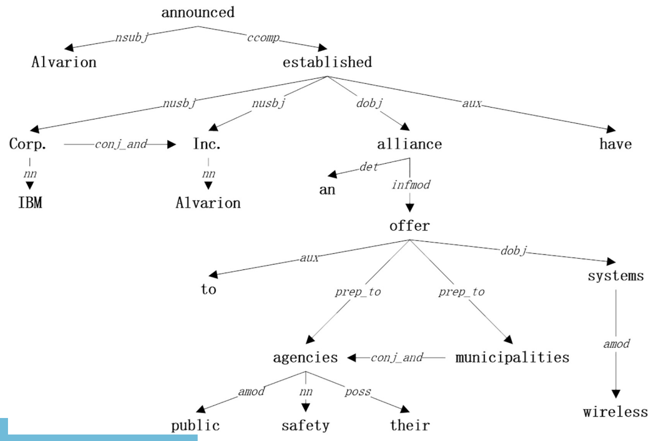


**Figure 3.**
Original dependency parse tree

tree captured their semantic meanings. This example used the Stanford Parse Tree (Manning *et al.*, 2014).

*4.3.2 Merging chunks.* The structure of an original parse tree is often too complex for alliance extraction template–matching. To reduce the template concept space, we used chunk parsing results from the previous step to simplify the tree structure by merging words into chunks. The following rules were used to merge keywords of the same phrases into chunks.

(1) If a dependency *d*(*w*1, *w*2) is within the dependency class MODIFIER and not one of RELATIVE_CLAUSE_MODIFIER, PURPOSE_CLAUSE_MODIFIER, PREPOSITIONAL_MODIFIER, ADV_CLAUSE_MODIFIER, TEMPORAL_MODIFIER, PRECONJUNCT, PARTICIPIAL_MODIFIER or INFINITIVAL_MODIFIER, merge w1 and w2 together to form a compound entity.

(2) If a dependency *d*(*w*1, *w*2) is within the dependency class AUX_MODIFIER and not COPULA, merge *w*1 and *w*2.

(3) If a dependency *d*(*w*1, *w*2) is within the dependency class PREPOSITION_MODIFIER, and w1 is not a verb, combine *w*1 and *w*2.

We kept the main word of a chunk as its head. We could classify all chunks into noun chunks or verb chunks. In the above example, seven chunks were extracted, as shown in Table 4. Figure 4 presents the simplified dependency parse tree. After chunking all related words, the grammatical structure of thes entence is simplified. Most of the remaining dependency relations belong to dependency classes SUBJECT, COMPLEMENT, and PREPOSITION.

*4.3.3 Alliance relation extraction.* Alliance relation extraction, a challenging task (Witten *et al.*, 2004), involves annotating the unstructured text with entities and entity relations. The methods used in extracting relations have two categories: learning-based and template-based approaches, the former of which was not suitable here because few alliance instances occur in news articles. Instead, we adopted *a template-based approach*, which relies on the identification of templates carefully crafted by experts to extract alliance relations. The identification of templates relies on (1) an initial expert domain lexicon, (2) an expanded lexicon, (3) expanded Noun Phrases (NPs) and Verb Phrases (VPs) and (4) identification of template. This process is illustrated in Figure 5. We listed the number of lexicon category, keywords, NPs and VPs, and templates in the figure as well.

(1) *Initial Lexicon.* To identify final templates, we start with a domain lexicon (Feldman *et al.*, 1998, 2002). Many existing research on relation extraction has used simple key phrase lists as domain lexicons (Bao *et al.*, 2008; Lau and Zhang, 2011).

One major challenge to identifying alliances is that many “informal partnerships” share similar keywords with formal alliances and are thus liable to incorrect identification as

| Chunks | Head |
|---|---|
| IBM Corp. | Corp. |
| Alvarion Inc. | Inc. |
| Have establish | Establish |
| An alliance | Alliance |
| To offer | Offer |
| Wireless systems | Systems |
| Their public safety agencies | agencies |

**Table 4.**
Extracted chunks from IBM example

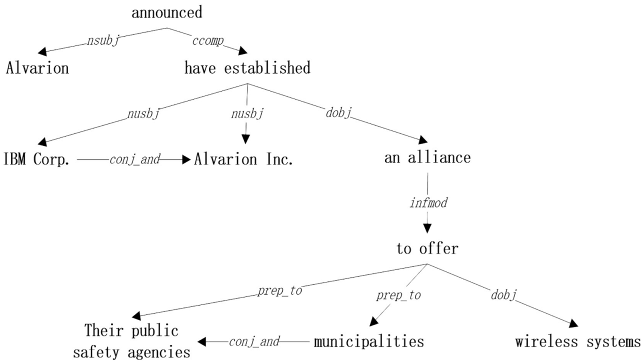


**Figure 4.**
Simplified dependency parse tree.

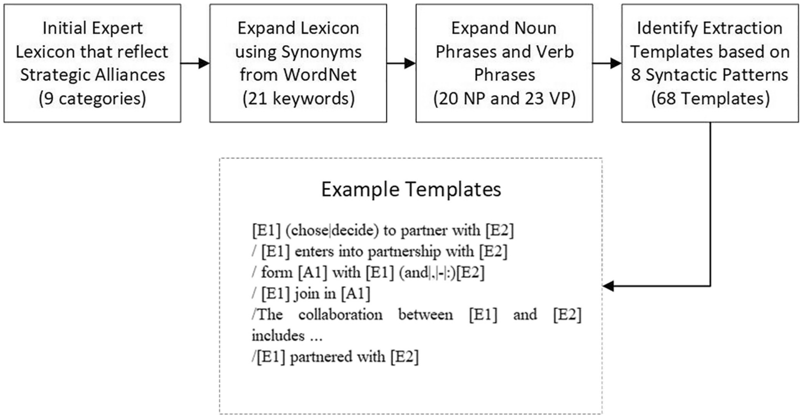

Initial Expert Lexicon that reflect Strategic Alliances (9 categories)
Expand Lexicon using Synonyms from WordNet (21 keywords)
Expand Noun Phrases and Verb Phrases (20 NP and 23 VP)
Identify Extraction Templates based on 8 Syntactic Patterns (68 Templates)
Example Templates
[E1] (chose|decide) to partner with [E2]
/ [E1] enters into partnership with [E2]
/ form [A1] with [E1] (and|,|-|:)[E2]
/ [E1] join in [A1]
/The collaboration between [E1] and [E2] includes ...
/[E1] partnered with [E2]

**Figure 5.**
Template identification process

strategic alliances. For example, customer and supplier relationships are partnerships, but not a formal alliance. While we included all the keywords indicative of both partnerships and alliances in our news search (crawling) phase, we had to differentiate between partnerships and formal alliances by assigning them different weights during the alliance extraction phase. Our domain experts further researched each keyword and decided their weights using a combination of their domain knowledge and evidence from news articles. Table 5 presents a sample of these keywords. Ambiguous keywords, such as “team with” and “work with”, considered weak keywords for alliance extraction, usually retrieve more news articles than more specific keywords. More specific keywords and phrases, such as “joint venture” and “join forces”, which retrieve fewer articles, are considered strong keywords. In our framework, we gave more power to strong keywords in extracting alliances.

(1) *Expand Lexicon.* To further expand our lexicon, we used WordNet (Miller, 1995) to identify the synonyms of alliance keywords and added these keywords to our lexicon. resulting with 21 keywords. In the classification phase, strong keywords, weak keywords, and WordNet keywords were assigned weights of 3, 2, and 1, respectively, to represent their extraction power.

(2) *Expand NPs VPs.* The next step is to expand the keywords into phrases, in particular Noun Phrases and Verb Phrases. The phrases is one step to reduce Using these four

| Expert identified keywords | Number of news extracted | Keyword frequency | Power |
|---|---|---|---|
| Partner (partnership/partnering/partnered) | 799 | 3,618 | Weak |
| Joint venture/initiative | 189 | 330 | Strong |
| Alliance | 183 | 904 | Strong |
| Collaboration (collaborate/collaborated) | 154 | 672 | Weak |
| Team with (teaming with/teamed with) | 1,602 | 1,989 | Weak |
| License (licensed/licensing) | 313 | 1,054 | Strong |
| Work together (worked together/working together) | 44 | 310 | Weak |
| Work with (worked with/working with) | 1,554 | 4,331 | Weak |
| Join forces (joined forces) | 10 | 55 | Strong |
| Total | 4,848 | 13,263 | |

**Table 5.** Expert identified initial keywords and their news retrieval rates

types of templates, together with our VPs and NPs, we finally identified a total of 68 templates.

(3) *Identify Templates.* Finally, the extraction of template was also domain-specific. Table 6 illustrates a general template of verb phrases. By using a well-defined, template-based extraction approach, Banko and Etzioni (2008) concluded that nearly 95% of 500 randomly selected sentences could be matched to one of the eight categories listed here.

Most current techniques are based on standard datasets and extract many types of relations not limited alliances. To make them work in a new domain, we studied both the syntactic and semantic sentence structures of alliance announcements. Following Banko and Etzioni's (2008) templates, we found four alliance announcement templates (Table 7) we call the *Alliance Discovery Template (ADT).*

Sothat the entity-extraction component would extract all organization names, we further modified the relation-extraction step so the templates only needed to contain one organization name. According to our template and a specialized lexicon list, we could predict the other organization name. With the results from the simplified dependency parse tree, we could implement the four modified templates in the following steps:

*Template 1.* If two dependencies $d1(w1, w2)$ and $d2(w3, w4)$ are within the dependency class SUBJECT and $w1$, $w3$ are the same chunk, we extract $w2$, $w4$, and $w1$ as our candidate alliance. This rule can be used to extract alliances that involve more than two organizations. (We can check if $w2$ and $w4$ contain organization names.)

<table>
<tr><th>Relative freq.</th><th>Category</th><th>Lexical syntactic pattern</th></tr>
<tr><td>37.8</td><td>Verb</td><td>E1 Verb E2 (e.g., X established Y)</td></tr>
<tr><td>22.8</td><td>Noun + Prep</td><td>E1 NP Prep E2 (e.g., X settlement with Y)</td></tr>
<tr><td>16.0</td><td>Verb + Prep</td><td>E1 Verb Prep E2 (e.g., X moved to Y)</td></tr>
<tr><td>9.4</td><td>Infinitive</td><td>E1 to Verb E2 (e.g., X plans to acquire Y)</td></tr>
<tr><td>5.2</td><td>Modifier</td><td>E1 Verb E2 Noun (e.g., X is Y winner)</td></tr>
<tr><td>1.8</td><td>Coordinaten</td><td>E1 (and|,|-|:) E2 NP (e.g., X-Y deal)</td></tr>
<tr><td>1.0</td><td>Coordinatev</td><td>E1 (and|,) E2 Verb (e.g., X, Y merge)</td></tr>
<tr><td>0.8</td><td>Appositive</td><td>E1 NP (:|,)? E2 (e.g., X hometown: Y)</td></tr>
</table>

**Table 6.** Verb phrase categories of template-based approach (Banko and Etzioni, 2008)

*Template 2.* If there are two dependencies *d*1 and *d*2, where *d*1(*w*1, *w*2) is within the dependency class SUBJECT and *d*2(*w*1, *w*3) is within the class COMPLEMENT, we extract *w*1, *w*2, and *w*3 as our candidate alliance.

*Template 3.* We extract two noun chunks connected by a dependency of PREPOSITION_between.

**Table 7.**
Alliance discovery template (ADT)

| | ADT | Example |
|---|---|---|
| Template 1 | Organization list + verb (form, establish, forge, etc.) | IBM, Sony and Toshiba form chip R&D alliance |
| Template 2 | First organization + verb (join, work with, etc.)+second organization | Red Hat joins top-level IBM strategic alliance |
| Template 3 | Noun (collaboration, agreement) + conjecture (between, among) +organization list | The collaboration between IBM and Geisinger . . . |
| Template 4 | Noun (participants, partners) + include + organization list | Participants include IBM and GE Health |

**Table 8.**
Features of ACRank

| Feature type | Features | Description | Confidence value |
|---|---|---|---|
| Sentence-level | Lexicon weight (LW) | Domain lexicon word (strong) | 3 |
| | | Domain lexicon word (weak) | 2 |
| | | WordNet associated word | 1 |
| | | None | 0 |
| | Entity extraction rate (EE) | ≥2entities extracted by multi-strategies | 3 |
| | | ≥2 entities extracted by single strategy | 2 |
| | | 1entity extracted by multi-strategies | 1 |
| | | 1entity extracted by single strategy | 0 |
| | Template used (*T*) | Template 3 | 4 |
| | | Template 4 | 3 |
| | | Template 1 | 2 |
| | | Template 2 | 1 |
| Document-level | Sentence position in document (*P1*) | Title | 3 |
| | | First paragraph | 2 |
| | | Last paragraph | 1 |
| | | Rest of the document | 0 |
| | Sentence position in paragraph (*P2*) | First sentence | 2 |
| | | Last sentence | 1 |
| | | Rest of the paragraph | 0 |
| | Number of entity co-occurrences in the entire document (CO) | More than 5 | 5 |
| | | 5 times | 4 |
| | | 4 times | 3 |
| | | 3 times | 2 |
| | | Twice | 1 |
| | | Once | 0 |

*Template 4.* If two dependencies $d1(w1, w2)$ and $d2(w1, w3)$ are within the dependency class COMPLEMENT and $d3(w1, w4)$ is within the class SUBJECT, we extract $w1$, $w2$, $w3$, and $w4$ as our candidate alliance. This rule can be used to extract alliances that involve more than two organizations. (We can check if $w2$ and $w3$ contain organization names.)

### *4.4 Corpus-level alliance ranking*

At the sentence level, we identified alliance instances using template-based relation extraction. However, this method generated many false positives. Therefore, we transcended template-based relation extraction and expanded from individual alliance instances to aggregating individual instances with a corpus-level Alliance Confidence Ranking algorithm called *ACRank*.

*4.4.1 Alliance feature set.* Each extracted relation instance takes on a different confidence level according to multiple features. In our design framework, we identified six important features (Table 8). The first three focus on the sentence-level view. Lexicon weight (LW) examines the appearance of critical words and phrases relevant to alliance, which come from the domain lexicon. These keywords are further classified into “strong” and “weak” indicators of alliance based on their power in finding true alliance instances. These lexicon keywords are further supplemented with their synonyms retrieved from WordNet to build an expanded lexicon with better coverage of alliance-related keywords. Extracted entity type (EE) examines the contribution of each strategy to the extraction of organization entities and the number of entities extracted in each sentence. Notice that our ADT only requires *one entity* to be extracted in a template and can predict the other entity or entities. However, the more entities a strategy can identify, the higher its confidence value is. Also, a strategy has a higher confidence value when it can extract entities that have also been extracted by many other strategies. Template used ($T$) is derived from our initial study of template performance. In our pilot study, we found that, regarding the number of accurate extractions, T3>T1>T4>T2, and we ranked the template confidence value in that order.

Document-level features include two positional features and the entity co-occurrence frequency feature. $P1$ captures the positions of extracted alliance instances in the entire article. The keywords in the title and the first paragraph of an article carry higher weight than those in other parts of the article. Similarly, $P2$ captures the positions of alliance instances in paragraphs. CO captures entity co-occurrences throughout the document. The more frequently two company names co-occur, the more likely they formed an alliance.

*4.4.2 ACRank.* After constructing the feature set of an extracted alliance instance, we looked at the appearance of the same alliance (containing the same organizations’ names) in the entire news article corpus. Multiple occurrences of alliance instances reinforce the confidences of relation extractions. Thus, we aggregated the confidence value of each alliance instance in the scope of the whole corpus and derived the final confidence value of an alliance as the Alliance Confidence Rank (*ACRank*):

$$ACRank\,Value = \sum_{i=1}^{n} \left( \frac{Confidence\,Value(i)}{n} \right) \tag{1}$$

where $n$ is the total number of occurrences of the same alliance in the extraction results.

*Confidence Value(n)* is derived from the six features described in Table 8 using formula 2:

$$Confidence\,Value(n) = \alpha LW(n) + \beta EE(n) + \gamma T(n) + \delta P1(n) + \varepsilon P2(n) + \zeta EO(n) \tag{2}$$

where $\alpha$–$\zeta$ are weights assigned to each feature.

In practice, we can empirically determine the optimal weights of $\alpha$–$\zeta$ by labeling another training dataset with a balanced number of alliance/non-alliance company relations. One possible approach is to follow a feature weighting scheme similar to that used by Pang *et al.* (2002) in sentiment classification. We can calculate the sum of feature values in the alliance instances and then divide it by the sum of feature values from the entire training dataset. We can also use an effective feature-weighting learning algorithm such as RELIEF (Kira and Rendell, 1992) or LFE (Sun and Wu, 2008), which can make estimations of parameter weights to optimize the classification performance. However, due to the space constraints of this paper and the high cost of constructing additional training datasets, we will not include the feature weight–optimization algorithms and simply give a weight value of 1 for all features in our case study.

If multiple alliance instances of the same organizations exist, *ConfidenceValue (1)* to *ConfidenceValue (n)* will be in descending order. We avoid a simple summation or average of *ConfidenceValue* because it would greatly favor large organizations whose names appear daily in the news. Capturing the alliances of mid-size and small organizations and foreign organizations is also important. Thus, we introduced a degrading factor $n$ as the denominator in Formula 1. However, the strongest evidence with the highest confidence value is not degraded.

We chose a ranking approach instead of a learning-based classification approach because alliances do not appear frequently in news articles. Our pilot study showed that the learning-based classification approach fails when the dataset is extremely skewed (i.e., few positive examples). A ranking approach also gives researchers flexibility to judge possible alliances later at their discretion. In our alliance-extraction results, even some rumored alliances might interest researchers because they might become true alliances in the future.

## 5. Case study and evaluation

Following our framework design, we developed an alliance-extraction system for the evaluation experiment. To evaluate the effectiveness of our proposed framework, we conducted a case study of alliances formed by IBM in 2006 because, as multinational and technology-heavy, IBM has established numerous alliances with many organizations inside and outside of the United States. The case study had three goals: (1) evaluate the effectiveness of our sentence-level ADT-based approach, (2) evaluate the effectiveness of our corpus-level ACRank approach, and (3)study the coverage of the Thomson Reuters SDC database by comparing its alliances with the alliances extracted by our system and domain experts. We chose Thomson Reuters SDC because it is the most popular commercial alliance database that records all publicly announced alliance deals globally. It identifies a wide range of strategic alliances which include R&D agreement, sales and marketing agreement, and supply agreement between multiple types of organizations which include business, government and universities based on the alliance mentioning made in the Securities and Exchange Commission (SEC) fillings, industry publications and news articles (Schilling, 2009).

### *5.1 Dataset*

An alliance meta-search lexicon was created by experts to search for relevant articles from multiple resources. The lexicon contained keywords such as “alliance,” “joint venture,” “team with,” “license,” etc. News articles were crawled from LexisNexis using this lexicon. LexisNexis provides a meta-search function to search for full-text news from highly reliable sources globally, including the world’s major newspapers, magazines, and trade publications. When we combined news from multiple sources, the duplicates were identified and removed from our news data collection. In our experiment, a total of 4,261 unique documents were crawled. Since it would be unrealistic to ask a human expert to read through all these articles, we randomly selected a subset of 1,000 news articles and asked an expert to manually read all the selected documents to define our human "gold standard."From this news dataset, the human expert identified a total of 63 true alliances.

Table 9 summarizes the statistics of our document collection and the extracted alliance instances after running the ADT template extraction. Notice that the ADT method does not require both organization entities to be extracted because the second entity can be predicted with template slots. There were 23 alliance instances in total with predicted second entities that would have been missed if the template had relied solely on EE. These unique alliances were then ranked.

### *5.2 Evaluation of ADT-based extraction*

We then compared our ADT-based extraction with two benchmark algorithms. Benchmark 1 used a co-occurrence–based approach, and benchmark 2 used a co-occurrence–based approach reinforced with critical verb phrases identified by experts. The former assumed that, if two organization names appeared in the same sentence, their relation was more likely an alliance, given that it was in an alliance-relevant news collection. We expected that the co-occurrence–based approach would achieve higher recall but lower precision because of the noise introduced. Therefore, we added a co-occurrence plus critical verb-based approach that extracted the possible alliances only when both organization names and at least a verb phrase from our domain lexicon appeared together.

Table 10 presents the performance of the ADT method compared to the two benchmark methods and the performance of each template extraction. The higher precision of the ADT method met our expectation because it utilized numerous linguistic parsing techniques, including POS tagging and dependency tree parsing. It also owed its higher recall to its ability to compensate EE by predicting the second entity in the template, while both co-occurrence–

| | | |
|---|---|---|
| Articles | Total number of articles | 1,000 |
| | Number of documents containing "IBM" | 985 |
| | Number of documents containing two organization entities (co-occurrence) | 642 |
| | Documents containing template instances | 314 |
| Sentences | Total number of sentences | 63,019 |
| | Number of sentences containing "IBM" | 3,289 |
| | Number of sentences containing two organization entities (co-occurrence) | 2,107 |
| Alliance instances | Total number of instances extracted by ADT | 329 |
| | Number of instances extracted by ADT containing two organization entities | 216 |
| | Number of instances extracted by ADT containing only one organization entity (with ADT predicting the other one) | 113 |
| Unique alliances | Number of unique alliances extracted by ADT | 126 |
| | Number of unique alliances extracted by ADT containing two organization entities | 103 |
| | Number of unique alliances extracted by ADT containing only one organization entity | 23 |

**Table 9.**
Summary of case study data and initial alliance extraction results

| | Method | Recall | Precision | *F*-measure |
|---|---|---|---|---|
| Lexicon only | Co-occurrence | 0.687 | 0.044 | 0.075 |
| | Co-occur + verb | 0.652 | 0.073 | 0.121 |
| Template-based | Template 1 | 0.558 | 0.511 | 0.528 |
| | Template 2 | 0.476 | 0.346 | 0.398 |
| | Template 3 | 0.114 | 0.571 | 0.186 |
| | Template 4 | 0.302 | 0.512 | 0.372 |
| | ADT | 0.781 | 0.447 | 0.568 |

**Table 10.**
Performance of the ADT method compared to benchmarks

based methods required two entities to be correctly identified in each sentence. Templates 1, 3, and 4 were better at finding alliances, Template 1 covering about 50% of the alliance announcements. We later used the precision performance of each template as their confidence levels in the ACRank component. In general, we found that lexicon-only method could identify most alliances. However, this method also generated a significant amount of false positives. ADT method provides a best balance between Recall and Precision with an F-measure of 0.568.

### *5.3 Evaluation of corpus-level ACRank*

Our second experiment evaluated the performance of the corpus-level ACRank by assigning equal weights to parameters $\alpha$–$\zeta$ in Formula 2 in Section 4.4.2. In terms of performance measures, we adopted the classic information-retrieval evaluation metrics of recall, precision, and *F*-measure for topN%–ranked documents. Formulas 3–5 show the calculation of these three measures from the alliance-extraction results. We changed the number of documents retrieved in the original functions to the number of alliances extracted to better fit our study.

$$P = \frac{\text{Number of correctly extracted alliances(instances)}}{\text{Number of all extracted alliances(instances)}} \tag{3}$$

$$R = \frac{\text{Number of correctly extracted alliances(instances)}}{\text{Total number of true alliances (instances)}} \tag{4}$$

$$F - \text{measure} = \frac{2PR}{P + R} \tag{5}$$

Table 11 presents the precision, recall, and *F*-measure comparisons of three methods: ACRank, sentence-level features-based alliance extraction, and document-level features-based alliance extraction with 10–100% top-ranked extracted alliances. We also added the extraction performance of an SVM classifier, a common machine-learning algorithm. SVM classification produced a precision rate of 59.3%, recall rate of 32.7% and *F*-measure of 0.422 with no percentage cutoff points. Figure 6 plots the precision/recall rates of these four extraction methods at different cutoff points from top10% to top100% of highly ranked alliances.

The recall rate consistently increased for each method as we included more alliances until the percentage of alliances reached 100%. In the ACRank results from the top50% of alliances, the recall rate is still the highest of all the methods at 73.5%. ACRank's precision rate also reached the highest point of 100% with the top10% of extracted alliances included and then gradually decreased to 44.7% when all alliances were included. With the top20% of alliance instances included, the precision rate of ACRank was maintained at 97%, a precision rate much higher than the rates of sentence-level/document-level features-based extraction and SVM, showing that our ACRank was most effective in predicting the very top alliances by having sentence-level and document-level features complement each other. The ACRank method maintained its precision rate at 65.5% when the top50% of alliance instances were included. Such performance is at a satisfactory level in comparison to many IE algorithms.

When looking at errors in our extraction results, we found three major causes of error: unidentified or misidentified entity names, incorrect dependency parse tree, and insufficient coverage of some alliance instances by our existing templates. These problems suggest that, although our ADT and ACRank achieved satisfactory performances, they have much room to improve. One way to increase both precision and recall rate is to use a more powerful EE algorithm and dependency parse tree. If we choose to add more templates to the model beyond these four, they may achieve a higher recall rate at the cost of a lower precision rate. We leave these questions for further investigation in future studies.

| | Top 10% | Top 20% | Top 30% | Top 40% | Top 50% | Top 60% | Top 70% | Top 80% | Top 90% | Top 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *Sentence-level* | | | | | |
| Precision | 72.7% | 60.6% | 57.6% | 54.5% | 54.5% | 52.8% | 50.9% | 49.0% | 46.3% | 44.7% |
| Recall | 16.3% | 27.2% | 38.8% | 49.0% | 61.2% | 70.7% | 79.6% | 87.8% | 93.2% | 100.0% |
| *F*-measure | 0.267 | 0.376 | 0.463 | 0.516 | 0.577 | 0.605 | 0.621 | 0.629 | 0.619 | 0.618 |
| | | | | | *Document-level* | | | | | |
| Precision | 66.7% | 60.6% | 53.5% | 50.0% | 47.3% | 47.2% | 46.1% | 46.4% | 46.6% | 44.7% |
| Recall | 15.0% | 27.2% | 36.1% | 44.9% | 53.1% | 63.3% | 72.1% | 83.0% | 93.9% | 100.0% |
| *F*-measure | 0.244 | 0.376 | 0.431 | 0.473 | 0.500 | 0.541 | 0.562 | 0.595 | 0.623 | 0.618 |
| | | | | | *ACRank* | | | | | |
| Precision | 100.0% | 97.0% | 81.8% | 69.7% | 65.5% | 60.4% | 53.9% | 50.6% | 47.6% | 44.7% |
| Recall | 22.4% | 43.5% | 55.1% | 62.6% | 73.5% | 81.0% | 84.4% | 90.5% | 95.9% | 100.0% |
| *F*-measure | 0.367 | 0.601 | 0.659 | 0.659 | 0.692 | 0.692 | 0.658 | 0.649 | 0.637 | 0.618 |
| | | | | | *SVM* | | | | | |
| Precision | | | | | 59.3% | | | | | |
| Recall | | | | | 32.7% | | | | | |
| *F*-measure | | | | | 0.422 | | | | | |

**Table 11.** Performance of ACRank with TopN% alliances

### *5.4 Comparison between ACRank and Thomson Reuters SDC*

In this experiment, we evaluated the precision and recall performance of our ACRank method in comparison with ane xisting alliance database: Thomson Reuters SDC. We selected it as our benchmark because most strategic alliance analyses are conducted manually using its data. The ACRank approach is designed to be a superior alternative alliance data source for researchers in addition to existing databases like Thomson Reuters SDC.

In the experiment, we compared the extracted alliances from our automated approach with the alliances available in the Thomson Reuters SDC database. Due to the enormous effort of data annotation by a human expert, the evaluation was done based on only1,000 documentsin our collection. From this dataset, experts identified 63 alliances. Table 12 presents the confusion matrix of the alliance extraction results from both methods, and Table 13 presents precision, recall, and *F*-measure.

From these results, we can see that though the Thomson Reuters SDC database made no identification mistakes, it only covered 7.9% of the total alliances identified by our experts from the news dataset. This is consistent with Schilling's (2009) conclusion that individual alliance data bases can only cover a subset of the whole population of strategic alliances. These comparison results demonstrate that the ACRank method can help build a much more comprehensive alliance database than Thomson Reuters SDC with a much higher recall rate (77.8%). However, in the ACRank extraction results, there were also 77 company relations which were not considered as strategic alliances by the experts. These false positive alliances significantly affected the precision rate of the ACRank. To further screen out the false positive alliances from the extraction results, alliance researchers and managers can verify the alliances, especially the alliances with lower confidence value by examining the news articles linked to the alliances available in our system. Balancing between precision and recall, ACRank method has better overall performance than Thomson Reuters SDC in strategic alliance extraction, as suggested by their *F*-measure values.

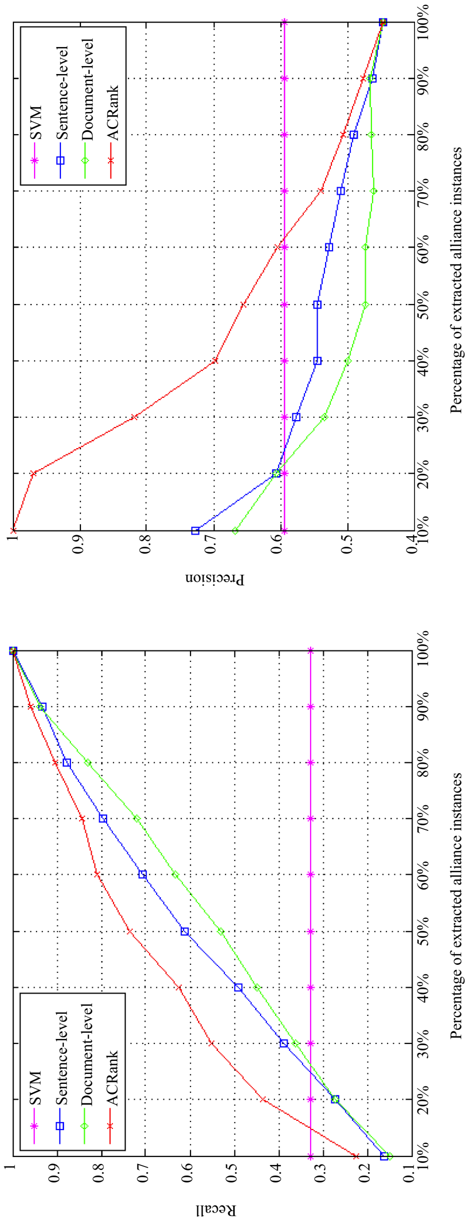


**Figure 6.**
Recall and precision of ACRank with TopN% alliances

### 5.5 Applying ACRank to Dow 30 companies

To further test the performance of ACRank, we expanded the data collection to include Dow Jones 30 publicly traded (Dow 30) companies, such as AT&T, Cisco System Co., and Boeing, with an expanded timeframe of 3 years (2006–2008). Using LexisNexis, we retrieved news articles about each company and identified alliance instances from each news article. We applied both ADT template extraction and ACRank on the collected news documents following the same procedure as described in Sections 4.3 and 4.4. We compared the top 100 alliances ranked by ACRank with the alliances available from the Thomson Reuters SDC database and calculated the recall rate based on how many percentages of Thomson Reuters SDC alliances could be identified using our approach for the top 100 ACRank results. To understand the coverage of ACRank of SDC, we used three measures:

$$\text{Worst} - \text{case Recall} = \frac{\text{ACRank Alliances} \cap \text{SDC Alliances}}{\text{SDC Alliances}} \tag{6}$$

$$\text{Best} - \text{case Recall} = \frac{\text{ACRank Alliances}}{\text{ACRank Alliances} \cup \text{SDC Alliances}} \tag{7}$$

$$\text{Overalapping} = \frac{\text{ACRank Alliances} \cap \text{SDC Alliances}}{\text{ACRank Alliances} \cup \text{SDC Alliances}} \tag{8}$$

In worst-case recall, we assume that only alliances identified by SDC are true alliances. In best-case recall, we used a pooling evaluation approach and assume that alliances identified by top-100 ACRank and SDC are all true alliances. This pooling approach is used in information retrieval when the set relevance documents may be impossible to be judged by human. The set relevance documents formed by creating relevance judgments for the pooled top *k* results of particular systems in a set of experiments (Aslam *et al.*, 2003, 2006). Since this measure is only based on 2 systems: ACRank and Thomson SDC Alliances. It represents the upper limit of the ACRank recall value, the name best-case measure was used here. We agree that the result is biased since Thomson often has less than *k* (100) results. We also report the overlapping ratio of ACRank and SDC.

Table 14 shows the number of news documents in our dataset, the number of Thomson Reuters SDC alliances, the number of overlapping Thomson Reuters SDC alliances in the top 100 ACRank results, worst-case recall, best-case recall and overlapping ratio for each company. While both the worst-case and best-case recalls are calculated based on the unlikely situations where the ACRank method either does not add new alliances into the SDC

| | Expert-approved | Expert-disapproved | | Expert-approved | Expert-disapproved |
|---|---|---|---|---|---|
| ACRank identified | 49 | 77 | Thomson SDC identified | 5 | 0 |
| ACRank missed | 14 | 0 | Thomson SDC missed | 58 | 0 |

**Table 12.** Confusion matrix of ACRank and Thomson Reuters SDC

| | Precision | Recall | *F*-measure |
|---|---|---|---|
| ACRank | 38.9% | 77.8% | 0.52 |
| Thomson Reuters SDC | 100% | 7.9% | 0.15 |

**Table 13.** Performance comparison of ACRank and Thomson Reuters SDC

**Table 14.** Coverage of alliances from ACRank top 100 and Thomson Reuters SDC results for Dow 30 companies

| Company name | Number of news articles | Thomson Reuters SDC alliances (2006–2008) | Overlapping ACRank alliances in SDC (from top 100) | Worst-case recall | Best-case recall | Overlapping |
|---|---|---|---|---|---|---|
| 3M Company (MMM) | 0 | 13 | 0 | N/A | N/A | N/A |
| Alcoa Inc. (AA) | 9,035 | 11 | 0 | 0 | 0.9 | 0 |
| American Express (AXP) | 9,461 | 14 | 2 | 0.14 | 0.89 | 0.02 |
| AT&T (T) | 19,728 | 11 | 5 | 0.45 | 0.94 | 0.05 |
| Bank of America BAC) | 31,105 | 8 | 3 | 0.38 | 0.95 | 0.03 |
| Boeing (BA) | 49,241 | 9 | 5 | 0.56 | 0.96 | 0.05 |
| Caterpillar (CAT) | 7,267 | 3 | 3 | 1 | 1.0 | 0.03 |
| Chevron (CVX) | 13,656 | 18 | 11 | 0.61 | 0.93 | 0.10 |
| Cisco System Co. (CSCO) | 10,097 | 19 | 19 | 1 | 1.0 | 0.19 |
| Coca-Cola Co. (KO) | 24,043 | 8 | 2 | 0.25 | 0.94 | 0.02 |
| DuPont de Nemous (DD) | 11,664 | 12 | 2 | 0.17 | 0.91 | 0.02 |
| Exxon Mobil (XOM) | 13,884 | 4 | 4 | 1 | 1.0 | 0.04 |
| Gamble Co. (PG) | 11,221 | 6 | 0 | 0 | 0.94 | 0 |
| General Electric Co. (GE) | 30,772 | 61 | 23 | 0.38 | 0.72 | 0.17 |
| Hewlett–Packard (HPQ) | 27,200 | 16 | 15 | 0.94 | 0.99 | 0.15 |
| Home Depot (HD) | 24,733 | 2 | 0 | 0 | 0.98 | 0 |
| Intel (INTC) | 17,624 | 28 | 27 | 0.96 | 0.99 | 0.27 |
| International Bus. Mach. (IBM) | 22,132 | 47 | 33 | 0.7 | 0.88 | 0.29 |
| J.P. Morgan chase (JPM) | 32,084 | 3 | 0 | 0 | 0.97 | 0 |
| Johnson and Johnson (JNJ) | 5,504 | 4 | 0 | 0 | 0.96 | 0 |
| Kraft Foods (KFT) | 2,549 | 3 | 2 | 0.67 | 0.99 | 0.02 |
| McDonalds Corp. (MCD) | 0 | 1 | 0 | N/A | N/A | N/A |
| Merck (MRK) | 11,399 | 23 | 8 | 0.35 | 0.87 | 0.07 |
| Microsoft Corp. (MSFT) | 59,975 | 95 | 87 | 0.92 | 0.93 | 0.81 |
| Pfizer Inc. (PFE) | 12,225 | 20 | 19 | 0.95 | 0.99 | 0.19 |
| Travelers Cos. Inc. (TRV) | 510 | 2 | 0 | 0 | 0.98 | 0 |
| United Technologies Corp. (UTX) | 1975 | 0 | 0 | NA | NA | NA |
| Verizon Communications Inc. (VZ) | 21,357 | 15 | 5 | 0.33 | 0.91 | 0.05 |
| Wal-Mart Stores | 27,504 | 2 | 1 | 0.5 | 0.99 | 0.01 |
| Walt Disney Co. (DIS) | 26,217 | 7 | 4 | 0.57 | 0.97 | 0.04 |
| Average |  |  |  | 0.48 | 0.95 | 0.10 |

knowledge base at all or makes no mistakes in alliance identification, these two values together give us references to the range where the true ACRank recall value lies. We observed that even in the worst-case recall, ACRank still achieved 0.48 in covering alliances available in the Thomson Reuters SDC database, excluding the companies about which we have either no news articles, or we have no records in the Thomson Reuters SDC database. In best-case scenario, the recall reaches 0.95. True recall of ACRank should be between these two numbers. While we observe overlapping alliances, the ratio is 10%. This again is consistent with our IBM case study and what researchers have estimated (Schilling, 2009). This further confirms that ACRank has the potential to extract additional strategic alliances to address the coverage issue with current alliance database. At the last, ACRank can be a tool for analyst to pre-screen potential alliances without the need to read all news articles.

## 6. Conclusions and implications

Strategic alliance information is increasingly important to economists, managers, and policymakers in their decision-making processes, which depend on access to information on opportunities and barriers for strategic alliances. Despite this rising interest, most researchers still rely on manually constructed, low-coverage databases to perform analysis and draw conclusions. Commercial databases, such as Thomson Reuters SDC, are the most popular sources for the study of strategic alliances (Basole *et al.*, 2015; Schilling and Phelps, 2007). Their alliance coverages, of course, are limited by the set of input documents which can be read by their analysts in designated time periods.

In this research, we designed, developed, and evaluated a text-mining framework to extract alliance knowledge from news. We combined many text-mining techniques, such as POS tagging, dependency tree parsing, template-based extraction, and alliance ranking, incorporating document-level and sentence-level features into its design. In the evaluation experiment, we found our model extracted alliance information better than many baseline methods, including the Thomson Reuters SDC alliance database and SVM classifier-based extraction, especially in terms of recall.

This research makes strong contributions and implications in both text mining and academic communities. On the text-mining side, our work addresses the challenging problem of extracting entity relations from extremely skewed, noisy datasets using a combination of ADT and ACRank approaches. We found a multi-strategy approach to improve strategic alliance extraction using experts' knowledge about domain lexicons and templates as well as various shallow and deep parsing techniques. The same extraction framework has potential applications to other areas, such as biomedical and social sciences, to extract knowledge of interest. On the side of economics study, our study shows great promise in automating the construction of a large alliance database with a more comprehensive coverage of strategic alliances than many alliance databases currently available. This database would provide critical information and rich evidence to economics and public-policy researchers. Though we focused on the extraction of formal strategic alliances in this work, other types of business relations can be extracted, such as product-customer partnerships and competitive relations, by extending our framework. All these relations combined can provide a "big picture" for business researchers to study the business ecosystem and networked organizations. In a broad sense, our work bridges the gap between information science and economics studies by using text-mining techniques to bolster the study of a topic of major interest to economics researchers. We hope this work fosters the awareness of cross-disciplinary research and inspires more collaborations between business, management, social science, and information science.

## 7. Future directions

We plan the following: further testing of our analytical framework by applying it to longitudinal news documents collected from more companies from many industry sectors, creating a much bigger news document dataset for testing to demonstrate the effectiveness of our approach in handling high volumes of data, further improving the performance of the ACRank algorithm by adding features and templates and tuning the confidence weight parameters for the sentence- and keyword-level features. We also plan to build an evidence-based alliance Web portal that allows researchers to link alliance records in the database with alliance announcements from the news to facilitate various types of in-depth analysis and visualization, such as frequency analysis, sector analysis, and trend analysis. With this portal, ACRank is a tool to alleviate extensive human effort, and to provide complimentary information that limited analyst may have neglected. Strategic decision maker, researchers may screen out the false positive alliances by examining the news articles they are linked to. This will assure a high coverage with much less manual effort, and a lower false positive rate during extraction.

## Notes

1. http://www.cis.upenn.edu/~treebank/.
2. http://opennlp.sourceforge.net/projects.html.
3. http://alias-i.com/lingpipe/.
4. https://opennlp.apache.org/.

## References

Aslam, J.A., Pavlu, V. and Savell, R. (2003), "A unified model for metasearch, pooling, and system evaluation", *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 484-491.

Aslam, J.A., Pavlu, V. and Yilmaz, E. (2006), "A statistical method for system evaluation using incomplete judgments", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development In Information Retrieval*, pp. 541-548.

Banko, M. and Etzioni, O. (2008), "The tradeoffs between open and traditional relation extraction", *Proceedings of the Association for Computational Linguistics (ACL) -08: Human Language Technology Conference (HLT)*, Columbus, OH, pp. 28-36.

Bao, S., Li, R., Yu, Y. and Cao, Y. (2008), "Competitor mining with the web", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20 No. 10, pp. 1297-1310.

Basole, R.C., Russell, M.G., Huhtamäki, J., Rubens, N., Still, K. and Park, H. (2015), "Understanding business ecosystem dynamics: a data-driven approach", *ACM Transactions on Management Information Systems (TMIS)*, Vol. 6 No. 2, p. 6.

Bunescu, R.C. and Mooney, R.J. (2005), "A shortest path dependency kernel for relation extraction", *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 724-731.

Chen, H., Chau, M. and Zeng, D. (2002), "CI Spider: a tool for competitive intelligence on the web", *Decision Support Systems*, Vol. 34 No. 1, pp. 1-17.

Culotta, A., McCallum, A. and Betz, J. (2006), "Integrating probabilistic extraction models and data mining to discover relations and patterns in text", *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 296-303.

D'Haen, J., Van den Poel, D., Thorleuchter, D. and Benoit, D.F. (2016), "Integrating expert knowledge and multilingual web crawling data in a lead qualification system", *Decision Support Systems*, Vol. 82, pp. 69-78.

Fan, W., Wallace, L., Rich, S. and Zhang, Z. (2006), "Tapping the power of text mining", *Communications of the ACM*, Vol. 49 No. 9, p. 77.

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y. and Zamir, O. (1998), "Text mining at the term level", *European Symposium on Principles of Data Mining and Knowledge Discovery*, pp. 65-73.

Feldman, R., Aumann, Y., Schler, J., Landau, D., Lipshtat, O. and Ben-Yehuda, Y. (2002), "Term-level text with mining with taxonomies", Google Patents.

Frank, A., Becker, M., Crysmann, B., Kiefer, B. and Schäfer, U. (2003), "Integrated shallow and deep parsing: TopP meets HPSG", *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 104-111.

Hagedoorn, J., Link, A.N. and Vonortas, N.S. (2000), "Research partnerships", *Research Policy*, Vol. 29 Nos 4-5, pp. 567-586.

Hall, B.H., Jaffe, A.B. and Trajtenberg, M. (2001), "The NBER patent citation data file: Lessons, insights and methodological tools", National Bureau of Economic Research.

Heimeriks, K.H. and Duysters, G. (2007), "Alliance capability as a mediator between experience and alliance performance: an empirical investigation into the alliance capability development process", *Journal of Management Studies*, Vol. 44 No. 1, pp. 25-49.

Joly, P.-B. and de Looze, M.-A. (1996), "An analysis of innovation strategies and industrial differentiation through patent applications: the case of plant biotechnology", *Research Policy*, Vol. 25 No. 7, pp. 1027-1046.

Kira, K. and Rendell, L.A. (1992), "A practical approach to feature selection", *Machine Learning Proceedings 1992*, Elsevier, Cambridge, MA, pp. 249-256.

Lafferty, J., McCallum, A. and Pereira, F.C. (2001), "Conditional random fields: probabilistic models for segmenting and labeling sequence data", *Proceedings of 18th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282-289.

Lau, R. and Zhang, W. (2011), "Semi-supervised statistical inference for business entities extraction and business relations discovery", *Proceedings of the 1st International Workshop on Entity-Oriented Search (EOS)*, pp. 41-46.

Lawrence, S. and Giles, C.L. (1999), "Accessibility of information on the web", *Nature*, Vol. 400 No. 6740, p. 107.

Manning, C.D., Manning, C.D. and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT press, Cambridge, MA.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. (2014), "The Stanford CoreNLP natural language processing toolkit", *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Mansouri, A., Affendey, L.S. and Mamat, A. (2008), "Named entity recognition approaches", *International Journal of Computer Science and Network Security*, Vol. 8 No. 2, pp. 339-344.

Miller, G.A. (1995), "WordNet: a lexical database for english", *Communications of the ACM*, Vol. 38 No. 11, pp. 39-41.

Miller, S., Fox, H., Ramshaw, L. and Weischedel, R. (2000), "A novel use of statistical parsing to extract information from text", *1st Meeting of the North American Chapter of the Association for Computational Linguistics.*

Molina, A. and Pla, F. (2002), "Shallow parsing using specialized hmms", *Journal of Machine Learning Research*, Vol. 2, pp. 595-613.

Mooney, R.J. and Bunescu, R. (2005), "Mining knowledge from text using information extraction", *ACM SIGKDD Explorations Newsletter*, Vol. 7 No. 1, pp. 3-10.

Nadeau, D. (2007), "Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision", PhD thesis, University of Ottawa.

Nahm, U.Y. and Mooney, R.J. (2002), "Text mining with information extraction", *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pp. 60-67.

Nooteboom, B. (1999), *Inter-firm Alliances: Analysis and Design*, Psychology Press, Hove, East Sussex.

Oxley, J. and Wada, T. (2009), "Alliance structure and the scope of knowledge transfer: evidence from US-Japan agreements", *Management Science*, Vol. 55 No. 4, pp. 635-649.

Pang, B., Lee, L. and Vaithyanathan, S. (2002), "Thumbs up?: sentiment classification using machine learning techniques", *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Vol. 10, pp. 79-86.

Ratnaparkhi, A. (1996), "A maximum entropy model for part-of-speech tagging", *Conference on Empirical Methods in Natural Language Processing*, pp. 133-142.

Schilling, M.A. (2009), "Understanding the alliance data", *Strategic Management Journal*, Vol. 30 No. 3, pp. 233-260.

Schilling, M.A. and Phelps, C.C. (2007), "Interfirm collaboration networks: the impact of large-scale network structure on firm innovation", *Management Science*, Vol. 53 No. 7, pp. 1113-1126.

Srihari, R.. (2000), "A hybrid approach for named entity and sub-type tagging", *Sixth Applied Natural Language Processing Conference*, pp. 247-254.

Stuart, T.E. (1998), "Network positions and propensities to collaborate: an investigation of strategic alliance formation in a high-technology industry", *Administrative Science Quarterly*, Vol. 43 No. 3, pp. 668-698.

Sun, Y. and Wu, D. (2008), "A RELIEF based feature extraction algorithm", *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 188-195.

Tan, A.-H. (1999), "Text mining: the state of the art and the challenges", *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, Vol. 8, pp. 65-70.

Tan, A.-H. and Lai, F.-L. (2000), "Text categorization, supervised learning, and domain knowledge integration", *Proceedings of the KDD-2000 International Workshop on Text Mining, Boston*, Vol. 20, pp. 113-114.

Tsuji, Y.S. (2002), "Organizational behavior in the R&D process based on patent analysis: strategic R&D management in a Japanese electronics firm", *Technovation*, Vol. 22 No. 7, pp. 417-425.

Tsuruoka, Y. (2006), "A simple C++ library for maximum entropy classification", PhD thesis, University of Tokyo.

Vonortas, N., Caloghiorou, Y. and Ioadbides, S. (2003), "Research joint ventures: a critical survey of theoretical and empirical literature", *Journal of Economic Surveys*, Vol. 17, p. 541.

Witten, I.H., Don, K.J., Dewsnip, M. and Tablan, V. (2004), "Text mining in a digital library", *International Journal on Digital Libraries*, Vol. 4 No. 1, pp. 56-59.

Yan, M., Yu, Y. and Dong, X. (2016), "Contributive roles of multilevel organizational learning for the evolution of organizational ambidexterity", *Information Technology and People*, Vol. 29 No. 3, pp. 647-667.

Yoshino, M.Y. and Rangan, U.S. (1996), "Strategic alliances: an entrepreneurial approach to globalization", *Long Range Planning*, Vol. 29 No. 6, pp. 909-910.

Zelenko, D., Aone, C. and Richardella, A. (2003), "Kernel methods for relation extraction", *Journal of Machine Learning Research*, Vol. 3, pp. 1083-1106.

**Corresponding author**
Yilu Zhou can be contacted at: yzhou62@fordham.edu