Acquiring Activities of People Engaged in Certain Occupations

Miho Matsunagi, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura

Tokyo Institute of Technology
matsunag@lr.pi.titech.ac.jp
{sasano,takamura,oku}@pi.titech.ac.jp

Abstract. We present a system to acquire knowledge on the *activities* of people engaged in certain occupations. While most of the previous studies acquire phrases related to the occupation, our system acquires pairs of a verb and one of its arguments, which we call activities. Our system acquires activities from sentences written by people engaged in the target occupations as well as from sentences whose subjects are the target occupations. Through experiments, we show that the activities collected from each resource have different characteristics and the system based on the two resources would perform robustly for various occupations.

Keywords: activity acquisition, occupational knowledge, social media

1 Introduction

Knowledge about people engaged in certain occupations is useful in many situations. For instance, it would be valuable for those who want to become medical doctors to know that doctors often read academic papers as well as perform surgery. As another example, e-commerce companies can recommend e-book readers to news reporters if they know that news reporters frequently take bullet trains and read books on the train.

There are several studies on acquiring knowledge about people with certain attributes (e.g., [11, 1]). These studies aim to extract phrases related to target attributes. For example, *movie* and *IMAX camera* are related to film directors if there are many sentences like

(1) The film director shot a movie with an IMAX camera.

However, such phrases do not always capture the characteristics of the target attribute. For example, film directors and actors are both related to *movie* but a film director *shoots* a movie, while an actor *appears in* a movie. This difference can be helpful for people looking for occupations related to *movie*. To capture such differences, we acquire usual *activities* related to the occupation, such as *shoot a movie*, where an activity is defined as a pair of a verb and one of its arguments except its subject.

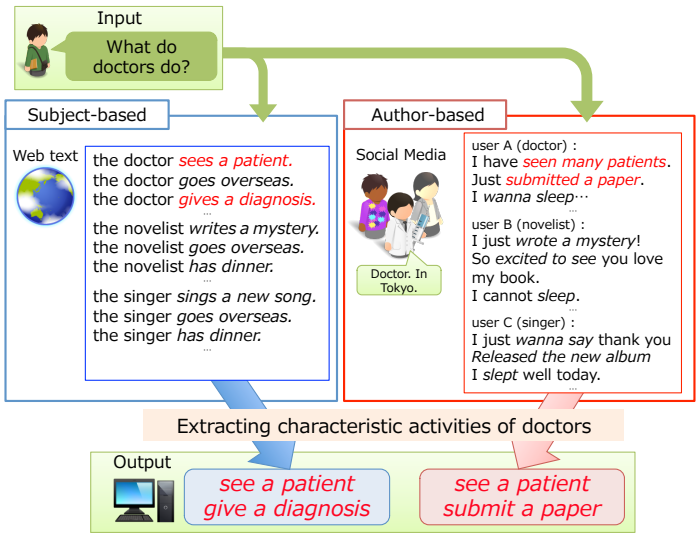

Fig. 1. Overview of our system.

In this paper, we present a system that acquires activities from two types of resources. The first type is sentences whose grammatical subjects are occupational titles. Example (1) is an instance of this type, from which we can extract *shoot a movie* and *shoot with an IMAX camera* as activities of film directors. This approach is based on the assumption that the grammatical subject of a sentence mostly denotes the agent of activities described in the sentence. Therefore, we can acquire activities related to the occupation by this approach. Most of these activities are, however, typical activities of the occupation, i.e. most of them are obviously associated with the occupation.

It is also valuable to know activities that are not obviously associated with the occupation but related to it as well as typical activities for those who are interested in the target occupations as their future occupation. For example, the medical doctor is known as an occupation that sees a patient and gives a diagnosis, but medical doctors often write and submit academic papers as researchers in their field. However, such activities as *submit a paper* cannot be collected from the first type of the sentences, because they rarely appear in these sentences. Thus, we also extract activities from sentences written in the first person by people engaged in target occupations, which we regard as the second type of sentences. If many medical doctors write "I just submitted a paper." on their blogs, we can extract *submit a paper* as an activity of medical doctors. This approach is based on the assumption that the sentences written in the first person often contain activities that rarely appear in the first type of sentences. We assume that activities with different characteristics can be collected by using both of these two resources.

Figure 1 shows an overview of our system. It consists of a subject-based component and an author-based component. The subject-based component collects

sentences of the first type mentioned above, i.e., sentences whose grammatical subjects are occupational titles, and extracts activities from them. The author-based component collects social media users engaged in target occupations and extracts these users' own activities from their posts.

2 Related Work

There are several studies on acquiring knowledge about certain attributes. Bergsma et al. [1] acquired knowledge on the properties of gender by calculating the pointwise mutual information between a gender and each phrase. Sap et al. [11] acquired knowledge on the properties of age and gender by using linear multivariate regression and classification models. These studies, however, focused on attributes with a limited number of classes, where it is easy to prepare labeled data and take a supervised approach. On the other hand, the number of occupations is much larger than that of age or gender and is usually not fixed. Therefore, it is impractical to prepare labeled data for all occupations in advance. We thus explore an unsupervised approach to acquire activities related to an input occupation.

The subject-based component of our system adopts a similar approach to unsupervised keyphrase extraction. Although most of the existing studies use various techniques such as language modeling [12] or graph-based ranking [13, 9], they are basically based on co-occurrence information between the target domain and keyphrases. We follow these studies in the subject-based component that uses co-occurrence information between the target occupation and an activity.

Our system also uses an author-based component that leverages social media text written by people engaged in target occupations. Although this component extracts authors' activities by looking for the first person, first person pronouns are often omitted in social media texts, especially those in Japanese, which is our focus. Kanouchi et al. [4] addressed a similar problem and built a supervised classifier to predict subjects for diseases/symptoms mentioned in sentences. On the other hand, we use several rules to extract authors' own activities in the first person sentences to remedy this problem.

Filatova and Prager [2] and Kozareva [7] also extracted activities from text. Filatova and Prager automatically extracted activities in the documents about a target person and classified them into occupation-specific or others. Kozareva acquired activities by which one could answer the question such as "What are the duties of a medical doctor?" for various entities, including persons, organizations, and other objects. However, these studies focus on only typical activities such as *see a patient* by doctors. On the other hand, we focus on not only typical activities but also non-typical activities such as *submit a paper* by doctors, and thus our target activities are more diverse than those of their studies.

3 Our System

As shown in Figure 1, our system acquires activities of an input occupation by using a subject-based component and an author-based component.

Table 1. Notation for frequencies regarding with a target activity and occupation for calculating the chi-square score. x and y in $N_{x,y}$ denote a target occupation and activity, respectively.

Type	Description
$N_{1,1}$	# of times that the target activity is performed by people with the target occupation
$N_{1,0}$	# of times that activities other than the target activity are performed by people with the target occupation
$N_{0,1}$	# of times that the target activity is performed by people with occupations other than the target occupation
$N_{0,0}$	# of times that activities other than the target activity are performed by people with occupations other than the target occupation

3.1 Subject-based Component

This component acquires activities from sentences whose grammatical subjects are occupational titles, because such sentences often contain activities of the target occupation. For example, both "*the doctor sees a patient*" and "*the doctor goes overseas*" in Figure 1 contain an activity of a medical doctor. Such activities, however, are not always specific to the target occupation. In Figure 1, while *see a patient* is specific to doctors, *go overseas* is not. Thus, we calculate the chi-square score χ^2 [10] between the activity and the target occupational title to measure how specific the activity is to the occupation.

Table 1 shows the notation related to frequencies $N_{i,j}$ for calculating χ^2. For example, $N_{0,1}$ of a doctor and *see a patient* denotes the frequency of *see a patient* by people other than doctors. χ^2 is calculated using $N_{i,j}$ by the following equations:

$$E_{i,j} = \sum_{i'} N_{i',j} \sum_{j'} N_{i,j'} / \sum_{i',j'} N_{i',j'},$$

$$\chi^2 = \sum_{i,j} (N_{i,j} - E_{i,j})^2 / E_{i,j}.$$

χ^2 compares the observed frequency of co-occurrence between the activity and the target occupation with the expected frequency of co-occurrence when the activity and target occupation are assumed to be independent of each other. We consider that one activity is related to the target occupation if its χ^2 score is large.

The process of this component is summarized as follows. It first collects pairs of a subject and an activity from parse trees. To avoid using incorrect parts of parse trees, it applies Kawahara and Kurohashi [5]'s method, which extracts unambiguous parts of the parse tree, and uses only reliable parts. It then calculates the χ^2 score between the input occupation and each activity and outputs activities with large χ^2 scores.

Table 2. Rules for extracting the authors' own activities. Only activities that satisfy these constraints are extracted.

Name	Description
Subject	The grammatical subject is the first person "I" or omitted.
Object	The grammatical object is not the author.
Modification	The verb representing the target activity does not modify a noun.
Modality	The modality of the sentence is not interrogatory, imperative, subjunctive, injunctive, or potential; these modalities suggest that the target activity might not actually be performed.

3.2 Author-based Component

This component acquires activities from social media texts in three steps. Since most social media posts are about the daily lives of users, we can collect activities related to the target occupation from their posts.

The component first collects users engaged in the target occupation from the profiles of social media users. Since some users describe their occupations in their profiles, we collect users whose profiles contain the target occupation. However, not all such users are actually engaged in the target occupation. Some users may mention an occupation that they want to have. Therefore, we use several rules to filter out users who are actually not engaged in the target occupation. When the target occupation is doctor, users with profile (2) are collected and users with profile (3) are not.

(2) I'm a doctor.

(3) My dream: to be a doctor.

Secondly, the component extracts users' activities from their posts. Since activities mentioned in users' posts are not always performed by the author of the post, we use several rules listed in Table 2 to select the authors' own activities[1]. As a result, the component extracts the underlined activities in (4) and (5), and filters out the activities in (6), (7) and (8).

(4) I just had dinner.

(5) Arrived at Tokyo.

(6) Call me when you get home.

(7) I saw a running dog.

(8) I should go to hospital.

In these examples, (6) is filtered because the object of call in (6) is the author, (7) is filtered because "running" in (7) modifies "dog.", and (8) is filtered because this sentence implies that "I" actually do not yet "go to hospital."

[1] The author-based component extracts the authors' own activities with the accuracy of 65.0% by using our rules. This accuracy does not directly affect the performance of the system, because specific activities are finally selected on the basis of χ^2 scores.

Table 3. Populations and accuracies of collecting users engaged in target occupations. Numbers in **bold** are less than 100 in population, and accuracies in *italic* are less than 60%.

Occupation	Acc.	(Pop.)	Occupation	Acc.	(Pop.)
announcer	69.0%	(206)	babysitter	84.5%	(2,819)
novelist	79.5%	(3,186)	homemaker	97.0%	(23,556)
photographer	85.5%	(1,664)	lawyer	90.0%	(428)
carpenter	*54.5%*	(625)	musician	90.5%	(694)
cook	66.5%	(367)	nurse	90.0%	(2,819)
counselor	91.5%	(618)	painter	88.3%	(552)
curator	89.6%	(**91**)	pharmacist	92.5%	(1,030)
detective	*11.1%*	(**9**)	pilot	14.0%	(290)
nutritionist	84.5%	(1,410)	civil-servant	80.0%	(1,502)
doctor	*59.0%*	(383)	singer	*60.0%*	(1,348)
editor	96.5%	(1,373)	station staff	*50.0%*	(**2**)
engineer	93.5%	(5,843)	teacher	71.0%	(1,664)
guard	*44.5%*	(1,528)	actor	88.5%	(239)
beautician	85.5%	(3,527)	news reporter	94.0%	(396)

Lastly, it calculates χ^2 scores in the same manner as in the subject-based component and outputs activities with large χ^2 scores as specific to the target occupation.

4 Experiment

4.1 Experimental Setting

We experimentally applied our system to the well-known occupations that satisfied all of the following constraints.

1. The occupation is registered as a noun in the dictionary of Japanese morphological analyzer JUMAN[2].
2. The occupation is listed in the Japanese Wikipedia's occupation list.
3. The occupational name appears more than 10,000 times in a Japanese Web corpus consisting of approximate 10 billion sentences.

Table 3 shows 28 occupations that were fed into our system. In the subject-based component, we used predicate-argument pairs extracted from approximately 6.5 billion parse trees[3]. In the author-based component, we crawled Twitter users who tweeted in Japanese in 2013 and their tweets by using Twitter API[4]. Consequently, we collected approximately 11,287,300 Japanese users. We finally extracted approximately 32,000 users tied to the target occupations

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[3] We used predicate-argument pairs provided by Kawahara and Kurohashi. For details of the method of extracting predicate-argument pairs from a Web corpus, please see Kawahara and Kurohashi [6].

[4] https://dev.twitter.com/overview/documentation

in the author-based component. We obtained syntactic structures from tweets by using a Japanese parser KNP [8].

As a preliminary experiment, we checked whether the author-based component correctly collected Twitter users engaged in the target occupation. Two human annotators examined 100 user profiles per occupation[5] and judged whether the automatically estimated occupations matched the occupations that the annotators considered the users have according to their profiles. Table 3 shows the population of collected users and accuracies. Although our author-based component failed to accurately collect users in some occupations that rarely appeared in social media, it collected the users with an accuracy of 80% for 22 occupations out of the 28 target occupations.

As mentioned in Section **??**, we assume that the activities collected from each resource have different characteristics. We thus conducted an evaluation with a crowdsourcing service to confirm this assumption. We presented an activity with a target occupation to the crowd-workers in the Japanese crowdsourcing service Lancers[6], and asked them to judge which of three categories the activity and occupation match.

1. **Obvious:** The presented activity is obviously associated with the presented occupation.
2. **Non-obvious:** The presented activity is not obviously associated with the presented occupation, but related to it in some way.
3. **Irrelevant:** The presented activity is not associated with the presented occupation.

In order to ensure the quality of the results, we selected only the workers who correctly answered quality control questions, which were very easy to answer if workers actually read the questions, such as *see a patient* by doctors. Each pair of an activity and an occupation was evaluated by five workers, and the score was calculated as the number of activities to which at least three out of five workers answered 1 or 2.

We evaluated occupations that had more than 200 activities in each component, because the quality of the acquired activities is not necessarily high if the number of acquired activities is too small. Both the subject-based and author-based components were evaluated for 13 out of 28 occupations, and only the author-based component was evaluated for 11 out of 28 occupations. Neither of these components were evaluated for remaining four occupations.

4.2 Comparisons of two types of activities

Table 4 shows the scores of each component. The subject-based component had accuracies of over 65% for 10 out of 13 evaluated occupations, though it failed in some occupations. In particular, the accuracies of homemakers and teachers

[5] Less than 100 users are collected for curator, detective, and station staff. Annotators examined all users for them.

[6] http://www.lancers.jp

Table 4. Scores of each component. The numbers in second and third columns denote the number of related activities out of 100 activities. For reference, we also show in the rightmost column the accuracies of collecting users by the author-based component in Table 3.

Occupation	The accuracy of activities acquisition (subject-based)	The accuracy of activities acquisition (author-based)	The accuracy of user collection
announcer	**74%**	56%	69.0%
novelist	**74%**	78%	79.5%
photographer	**85%**	95%	85.5%
cook	**91%**	48%	66.5%
counselor	**75%**	53%	91.5%
doctor	**66%**	19%	59.0%
engineer	**78%**	84%	93.5%
homemaker	33%	92%	97.0%
lawyer	**70%**	64%	90.0%
nurse	**76%**	78%	90.0%
singer	**85%**	81%	60.0%
teacher	38%	81%	71.0%
news reporter	59%	55%	94.0%
carpenter	-	17%	54.5%
nutritionist	-	36%	84.5%
editor	-	85%	96.5%
guard	-	10%	44.5%
beautician	-	68%	85.5%
babysitter	-	77%	84.5%
musician	-	95%	90.5%
painter	-	96%	88.3%
pharmacist	-	84%	92.5%
civil-servant	-	6%	80.0%
actor	-	63%	88.5%
Average	69.5%	63.4%	-

were lower than 40%. We manually examined the acquired activities of these occupations and found that the system often contain irrelevant activities that frequently appeared in advertisements on the Web. Most of these activities, such as *earn with a blog* by housemakers, do not match their real lives, and thus the accuracies for these occupations were low.

In the author-based component, the accuracies of collecting users engaged in the target occupations were correlated with the accuracies of acquiring activities in most cases. The component failed to acquire activities related to civil-servants and nutritionists, though their accuracies for collecting users were high, whose accuracies were 6% and 36%, respectively, because people with those occupations hardly wrote posts about their occupational activities. On the other hand, it successfully acquired activities related to singers though their accuracy of collecting

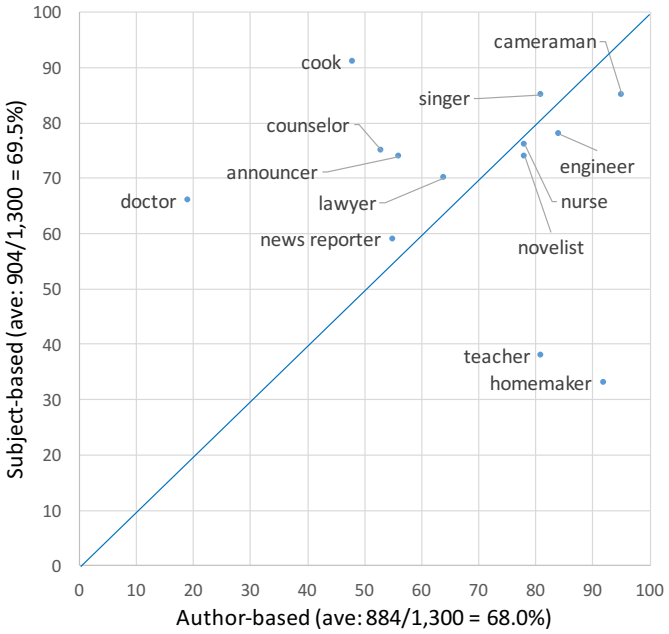

Fig. 2. Comparison of activities acquired by the two components. The x axis denotes the score of a author-based component, and the y axis denotes the score of a subject-based component.

users was not high, because they frequently announced the activities related to their occupations in the social media.

Next, we compared the activities for the 13 occupations, for which the two components were evaluated. Figure 2 compares the two components' accuracies in the scatter plot for 13 occupations. From Figure 2, we can see that both of the two components acquired activities for 13 occupations with the average accuracies of approximately 70%. However, the intersection of activities acquired by each component is very small; the average number of the common activities in the two components for 13 occupations is only 2.92 out of 100. This fact suggests that the system can acquire diverse activities by using both of the two components.

We then turn to the performance for each occupation. We found that the two components compensated for each other's weaknesses. For example, the subject-based component succeeded in acquiring activities of doctors and cooks, though the author-based component failed for them. Likewise, the author-based component succeeded in acquiring activities of homemakers and teachers, though the subject-based component failed for them. As a result, our system achieved an accuracy of at least 59% for each occupation by one of the components. Therefore, the system would become robust by combining the two components.

Table 5. The characteristics of correctly acquired activities. N_{total} is the sum of N_{ob} (obvious), $N_{non\text{-}ob}$ (non-obvious), and N_{other} (other) and it corresponds to the total score in Figure 2.

Component	N_{total}	N_{ob}	$N_{non\text{-}ob}$	N_{other}	$N_{non\text{-}ob}$ / ($N_{ob} + N_{non\text{-}ob}$)
Author-based	884	491	344	49	**41.2%**
Subject-based	904	577	281	46	32.8%

We further investigated the characteristics of the activities that were correctly acquired by the two components. As described above, we regarded the activities as correct if they were evaluated as "obvious" or "non-obvious" by at least three out of five crowd-workers. These activities were classified into the following three groups on the basis of the breakdown of the evaluation.

1. **Obvious:** The number of workers who evaluated the activity as "obvious" is larger than the number of workers who evaluated the activity as "non-obvious".
2. **Non-obvious:** The number of workers who evaluated the activity as "non-obvious" is larger than the number of workers who evaluated the activity as "obvious".
3. **Other:** The number of workers who evaluated the activity as "non-obvious" is equal to the number of workers who evaluated the activity as "obvious".

Table 5 shows the classified result. In this table, N_{total} denotes the total number of correctly acquired activities, and N_{ob} (obvious), $N_{non\text{-}ob}$ (non-obvious), and N_{other} (other) denote the number of activities the respective groups. The author-based component acquired more "non-obvious" activities than the subject-based component did. This difference was significant according to Fisher's exact test [3] at a significance level 0.01. This result supports our assumption that the activities collected from each resource have different characteristics.

We also investigated the relation between χ^2 rank and an accuracy. We first sort activities in descending order of χ^2 scores, and calculated the accuracy of top-N ($N = 10, 20, ..., 100$) activities. Figure 3 shows the result. From this figure, we found that the accuracy did not change regardless of χ^2 scores for the subject-based component, while the accuracy decreased monotonically in accordance with the increase of the activities for the author-based component. We think it is because the subject-based component occasionally gives high χ^2 scores to peculiar activities that were performed by only few people engaged in the target occupations. These activities often appear as phrases such as those in book titles and headlines, and thus they appear more frequently than the actual situations. However, note that when we manually investigated the activities with quite low χ^2 scores in the subject-based component, these activities were hardly related to the target occupations. Therefore, χ^2 score is effective for filtering out unrelated activities, while it fails to filter out some peculiar activities.

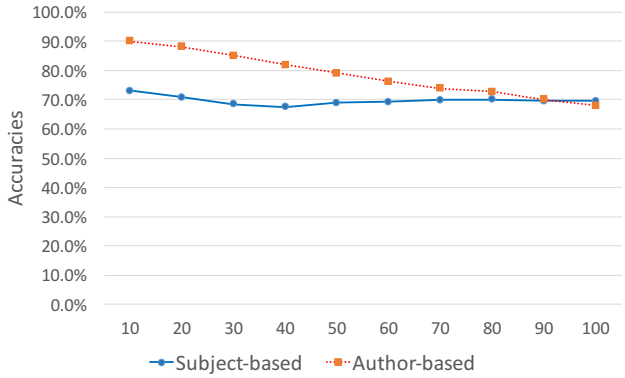

Fig. 3. The change of accuracies associated with the number of activities.

Table 6. Examples of activities acquired in each component. Numbers in brackets denote the number of crowd-workers who evaluated the activity as (obvious, non-obvious, irrelevant), respectively.

Occupation	Subject-based	Author-based
lawyer	accept a consultation (3,2,0)	write a brief (0,5,0)
	be in charge of the defense (5,0,0)	recruit a lawyer (2,2,1)
	establish a defense counsel (5,0,0)	appear in the office (0,4,1)
doctor	deceive a patient (2,1,2)	go to an academic meeting (1,4,0)
	charge for a treatment (5,0,0)	look down (0,1,4)
	write a medical certificate (5,0,0)	pay tax (0,0,5)
homemaker	succeed in business (0,0,5)	hang out the laundry (5,0,0)
	earn with a blog (1,2,2)	prepare a lunch box (5,0,0)
	try for pocket money (2,1,2)	take a daughter out (3,2,0)

Table 6 shows some examples from our system. The author-based component often acquired activities that were related to the target occupation and performed in the daily life, though it sometimes acquired activities that were not specific to the target occupation. For example, *prepare a brief* by lawyers is related to them and is likely to be performed in their daily lives, while *look down* and *pay tax* by doctors are performed by everyone. On the other hand, the subject-based component often acquired activities that were frequently mentioned by others, though it sometimes acquired peculiar activities that were not actually performed by the target occupation. For example, *accept a consultation* and *establish a defense counsel* are indeed typical activities of lawyers, while *deceive a patient* by doctors is actually unlikely to be performed by them. We think the subject-based component incorrectly acquired *deceive a patient* because it is often mentioned in the book titles and headlines as a doctors' activity.

5 Conclusion

We presented a system that had two components to acquire knowledge about the activities of people engaged in certain occupations. The subject-based component acquires activities from the sentences whose grammatical subjects are the target occupational title, while the author-based component acquires activities from text written by people engaged in the target occupation. In the evaluation with a crowdsourcing service, the subject-based component and author-based component acquired activities for 13 occupations with the average accuracies of 69.5% and 68.0%, respectively. As a whole, our system achieved an accuracy of at least 59% for each occupation by one of the components. We also showed that the activities acquired with each component have different characteristics. The author-based component acquired more activities that were not obviously associated with the occupation but related to it than the subject-based component did. For future work, we plan to explore the strategy for combining the two components that robustly acquires activities for various occupations.

Acknowledgement

We would like to acknowledge Prof. Kurohashi and Prof. Kawahara for providing us with the data of predicate-argument pairs used in the experiments. This work was supported by JSPS KAKENHI Grant Number JP26280080 and the Center of Innovation Program from Japan Science and Technology Agency, JST.

References

1. Bergsma, S., Van Durme, B.: Using Conceptual Class Attributes to Characterize Social Media Users. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). pp. 710–720. Association for Computational Linguistics, Sofia, Bulgaria (August 2013)
2. Filatova, E., Prager, J.: Tell Me What You Do and I'll Tell You What You Are: Learning Occupation-related Activities for Biographies. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP). pp. 113–120. Association for Computational Linguistics, Stroudsburg, PA, USA (2005), http://dx.doi.org/10.3115/1220575.1220590
3. Fisher, R.A.: On the interpretation of χ 2 from contingency tables, and the calculation of p. Journal of the Royal Statistical Society 85(1), 87–94 (1922)
4. Kanouchi, S., Komachi, M., Okazaki, N., Aramaki, E., Ishikawa, H.: Who caught a cold ? - Identifying the subject of a symptom. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP). pp. 1660–1670. Association for Computational Linguistics, Beijing, China (July 2015)
5. Kawahara, D., Kurohashi, S.: Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis. In: Proceedings of the 19th International Conference on Computational linguistics (COLING). pp. 425–431. Association for Computational Linguistics, Taipei, Taiwan (2002)

6. Kawahara, D., Kurohashi, S.: Case Frame Compilation from the Web using High-Performance Computing. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). pp. 1344–1347. Ganoa, Italy (2006)
7. Kozareva, Z.: Learning Verbs on the Fly. In: 24th International Conference on Computational Linguistics (COLING). pp. 599–610. The COLING 2012 Organizing Committee, Mumbai, India (December 2012)
8. Kurohashi, S., Nagao, M.: A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. Computational Linguistics 20(4), 507–534 (1994)
9. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: Lin, D., Wu, D. (eds.) Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (July 2004)
10. Miller, R., Siegmund, D.: Maximally selected chi square statistics. Biometrics 38(4), 1011–1016 (1982), http://www.jstor.org/stable/2529881
11. Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., Schwartz, H.A.: Developing Age and Gender Predictive Lexica over Social Media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1146–1151. Association for Computational Linguistics, Doha, Qatar (October 2014)
12. Tomokiyo, T., Hurst, M.: A Language Model Approach to Keyphrase Extraction. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment. pp. 33–40. Association for Computational Linguistics, Stroudsburg, PA, USA (2003)
13. Zha, H.: Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 113–120. SIGIR '02, ACM, New York, NY, USA (2002)