Trend Survey on Japanese Natural Language Processing Studies over the Last Decade

Masaki Murata[†], **Koji Ichii**[‡], **Qing Ma**[†,§], **Tamotsu Shirado**[†], **Toshiyuki Kanamaru**[†,*], **and Hitoshi Isahara**[†]
[†]National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{murata,qma,shirado,kanamaru,isahara}@nict.go.jp
[‡]Port and Airport Research Institute
Nagase 3-1-1, Yokosuka, Kanagawa 239-0826, Japan, ichii@pari.go.jp
[§]Ryukoku University, Otsu 520-2194, Japan, qma@math.ryukoku.ac.jp
[*]Kyoto University, Yoshida-Nihonmatsu, Sakyo, Kyoto 606-8501, Japan
kanamaru@hi.h.kyoto-u.ac.jp

Abstract

Using natural language processing, we carried out a trend survey on Japanese natural language processing studies that have been done over the last ten years. We determined the changes in the number of papers published for each research organization and on each research area as well as the relationship between research organizations and research areas. This paper is useful for both recognizing trends in Japanese NLP and constructing a method of supporting trend surveys using NLP.

1 Introduction

We conducted a trend survey on Japanese natural language processing studies that have been done over the last ten years. We used bibliographic information from journal papers and annual conference papers of the Association for Natural Language Processing, Japan (The Association for Natural Language Processing, 1995-2004; The Association for Natural Language Processing, 1994-2003). Just ten years have passed since the association was established. Therefore, we can use the bibliographic information from the past ten years. In this study, we investigated what kinds of studies have been presented in journal papers and annual conference papers on the Association for Natural Language Processing, Japan. We first digitized documents listed in the bibliographic information and then extracted various pieces of useful information for the trend survey.

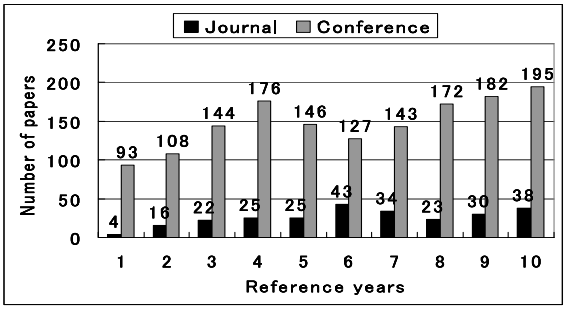

Figure 1: Change in the number of papers

We also examined the changes in the number of papers put up by each Japanese research organization and the changes in the number of papers written on specific research areas. Moreover, we examined the relationship between each Japanese research organization and each research area. This study is useful for trend surveys of studies performed by members of in the Association for Natural Language Processing, Japan.

2 Trend survey on NLP research studies

We show the changes in the number of journal papers and conference papers in Figure 1. Journal papers are reviewed, but conference papers are not reviewed in the association. In comparing the journal papers and conference papers, we found that the number of conference papers was much larger than that of journal papers. We also found that although both types of papers decreased in number at some point, they both demonstrate an upward trend.

Conference papers have a temporal peak in the fourth year and a temporal drop in the sixth year,

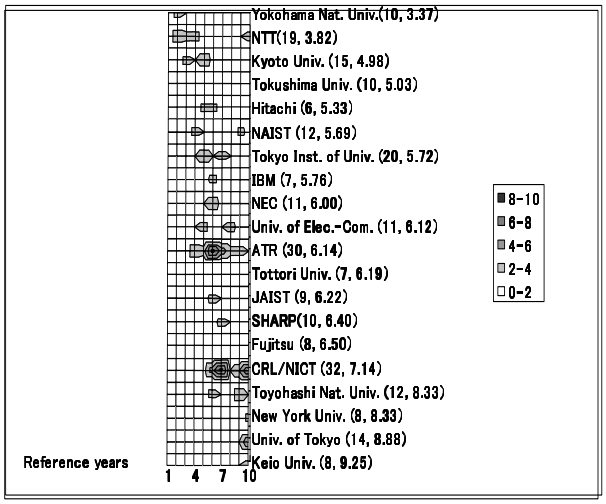

Figure 2: Change in the number of journal papers by each research organization (The two numbers in the parentheses indicate the total number of papers and the average value of published years.)

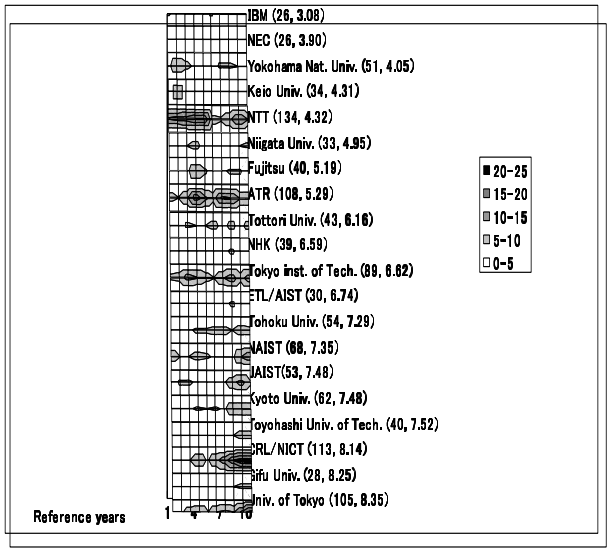

Figure 3: Change in the number of conference papers by each research organization

while journal papers have a peak in the sixth year and a drop in the eighth year. The temporal peak and drop of the journal papers occurred just two years after the peak and drop of the conference papers. We presume this is because journal papers need more time for reviewing and publishing, and because journal papers are presented later than conference papers for studies performed at the same time.

3 Trend survey on research organizations

Next, we investigated the change in the number of papers put out by each research organization. The results are represented in contour in Figures 2 and 3. The height in contour (the depth of a black color) indicates the number of papers. We calculated the average (we call it *average value*) of the average, the mode, and the median of the published years by using the data of the number of papers performed by each research organization. In the figures, each research organization is listed in ascending order of the average value. We added the total number of papers and the average value to each research organization in the figures. Therefore, research organizations that had many papers in the earlier years are displayed higher on the list, while research organizations that had many papers in the later years are displayed lower. Here, we displayed only research organizations that had many total papers. If a research organization's name was changed during the ten-year period, we used the name that had the most usage on published papers for displaying it.[1]

From these figures, we can see that ATR and CRL (NICT) put out many journal papers, and NTT, ATR, Tokyo Institute of Technology, CRL, and the University of Tokyo put out many conference papers. We also found that while NTT and ATR had many papers in the earlier years, CRL and the Univ. of Tokyo had many papers in the later years. We can expect that because CRL and the Univ. of Tokyo demonstrate an upward tendency, their quantity of papers will continue to increase in the future. Using these figures, we can see very easily in which reference year each research organization put out many papers.

4 Trend survey on research areas

Next, we investigated the change in the number of papers in each research area. The results are in Figures 4, 5, and 6. (Because the volume of data for conference papers was large, it was divided into two figures.). For journal papers, the height

[1]When we counted the frequency of a research organization whose name was changed, we used all the names of it including old and new names.

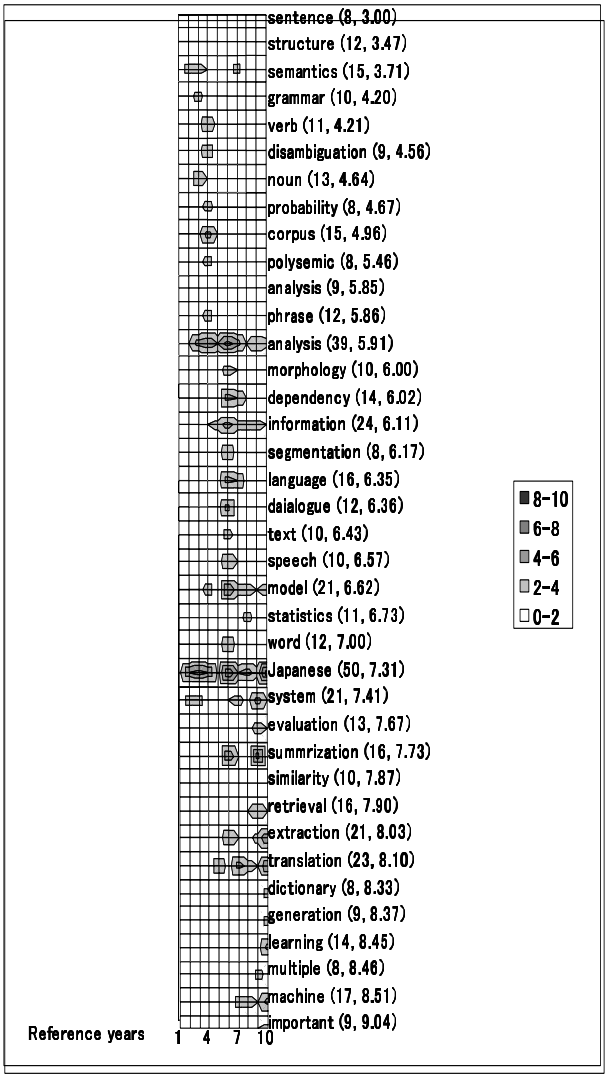

Figure 4: Change in the number of journal papers in each research area

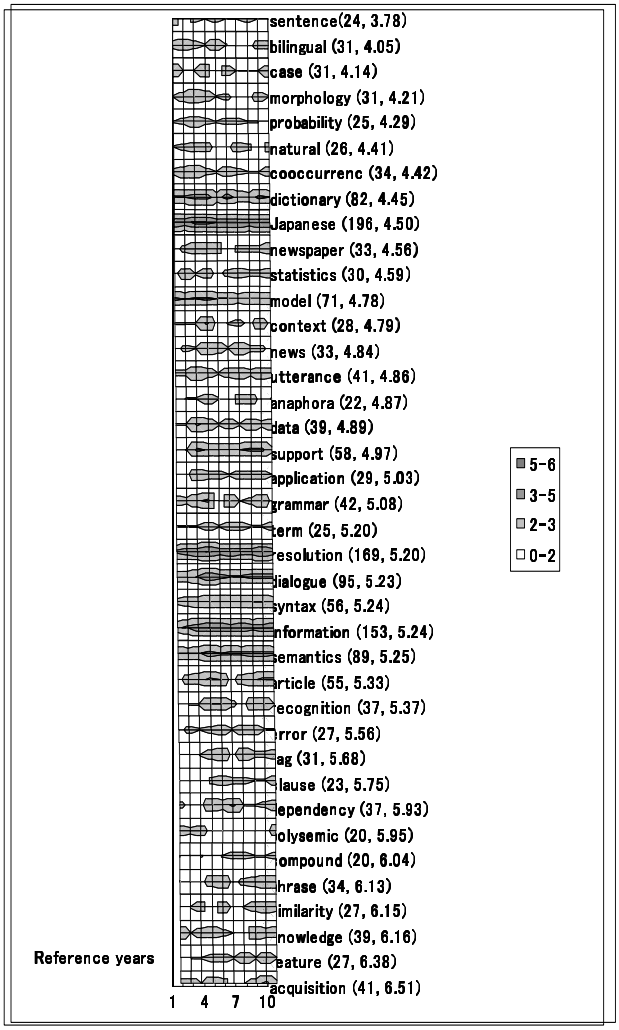

Figure 5: Change in the number of conference papers in each research area (part I)

in contour indicates the number of papers. For conference papers, the height in contour indicates the base two logarithm of the number of papers added by one. Using the same method as that described above, we calculated the average of the average, mode, and median of the years papers were published using the data of the number of papers in each research area. In the figures, each research area is displayed in ascending order of the average value. We added the total number of papers and the average value to each research area in the figures. Here, we divided the title of each paper into words using ChaSen software (Matsumoto et al., 1999), and we treaded each word as a research area. A paper with a particular word in its title was categorized in the research area indicated by the word. We manually eliminated words that were not indicative of a research area, for example, “teki” (of) and “kenkyu” (study).

From these figures, it is clear that the research areas of “Japanese” and “analysis” were studied in an especially large number of papers. We also found that for journal papers, because the research areas of “verb”, “noun”, “disambiguation”, “probability”, “corpus”, and “polysemic” were displayed higher on the list, these areas were studied thoroughly in the earlier years. Likewise, we found that the research areas of “morphology”, “dependency”, “dialogue”, and “speech” were studied thoroughly in the sixth year and the

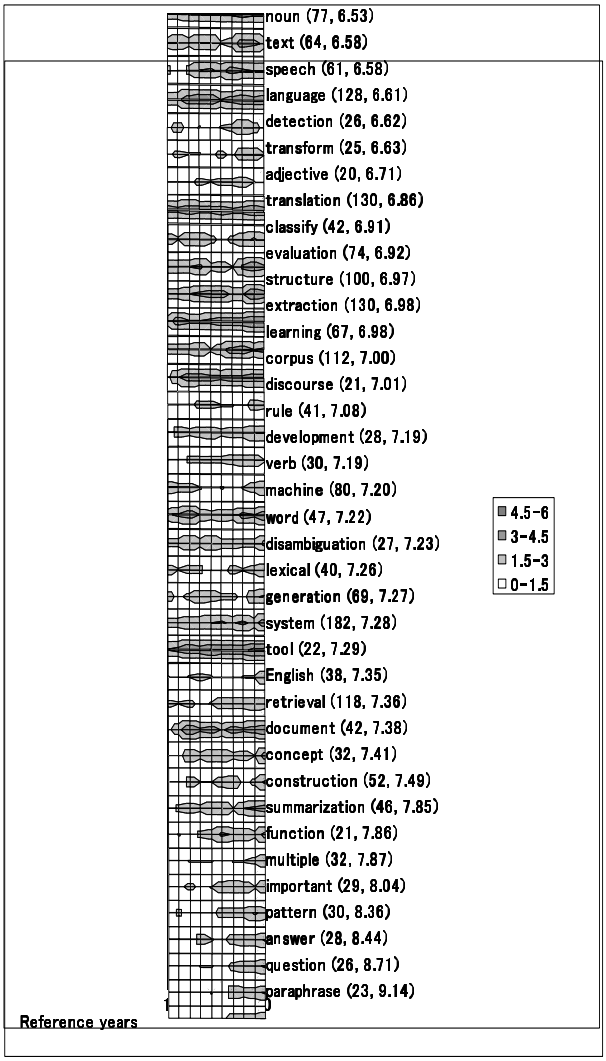

Figure 6: Change in the number of conference papers in each research area (part II)

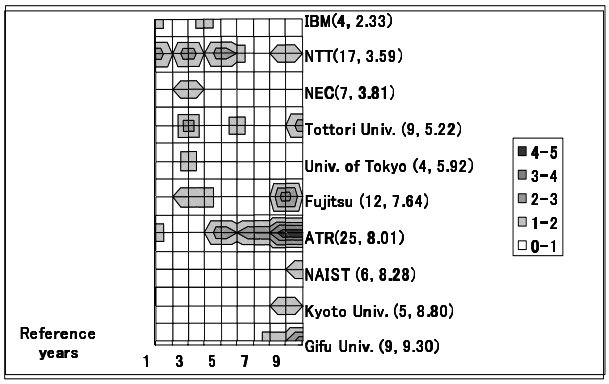

Figure 7: Change in the number of conference papers at each research organization in the research area of "translation"

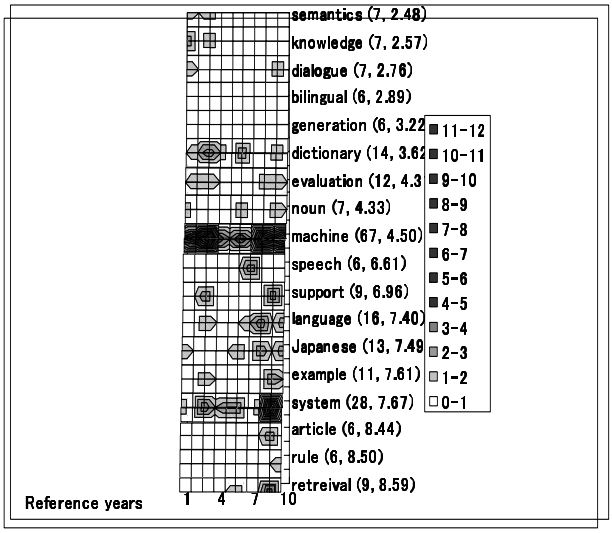

Figure 8: Change in the number of conference papers in each research area in the research area of "translation"

research areas of "summarization", "retrieval", "translation" and so on were studied well in the later years. Special journal issues on "summarization" were published in the sixth and ninth years, so the research area of "summarization" was represented in many papers in those years. We can expect that because the research area of "translation" demonstrates an upward tendency, the number of papers on this topic will continue to increase in the future.

In terms of conference papers, we found that the research areas of "bilingual", "morphology", "probability", "dictionary", "statistics", and so on were studied well in the earlier years. In the lower part of the figures, such research areas as "retrieval", "summarization", "question" and "paraphrase" are found. Thus, we can see that these research areas were studied thoroughly in recent years. We can see very easily in which reference years each research area was studied using these figures.

5 Trend survey using part of data

Although we have focused on using all the data in the trend survey so far, we can narrow down the survey by looking only at a certain part of the data. For example, when we want to exam-

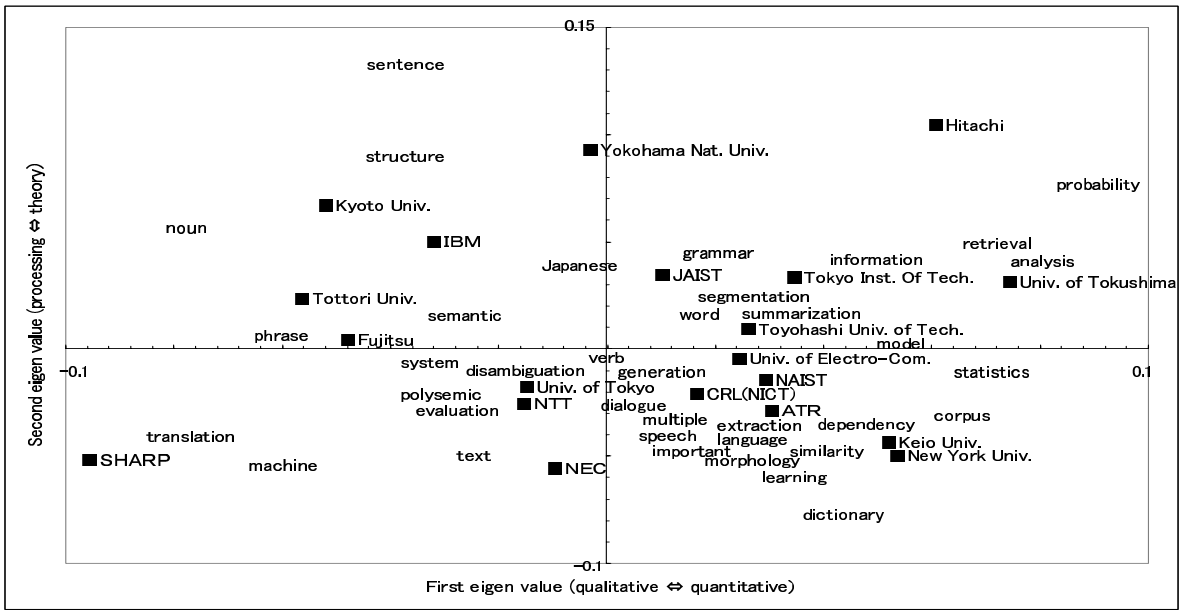

Figure 9: Relationship between research organizations and research areas in journal papers (The name of each research organization is given a "■" symbol.)

ine a trend survey on translation in more detail, all we have to do is to extract papers on translation and use them for a trend survey. We carried out a trend survey on machine translation in this manner. We first extracted papers whose titles included the word "translation" and then performed the same investigations as in Sections 3 and 4.

The results are in Figures 7 and 8. The height in contour (the depth of a color) indicates the number of papers. From Figure 7, we can see that NTT had many papers in the earlier years, and ATR had many papers in later years. From Figure 7, we can also see that studies on translation often dealt with specific topics such as "semantics", "knowledge" and "dictionary" in earlier years and "support", "example", and "retrieval" in more recent years.

6 Relationship between research organizations and research topics

Finally, we investigated the various research areas that research organizations studied more frequently during the ten-year period. Here, we show only the results for journal papers. We used the same method as in the previous sections for extracting research organizations and research areas from the data. We counted the cooccurrent frequency of each research organization and each research area. We then constructed a cross table in this manner and then performed the dual scaling method (Weller and Romney, 1990; Ueda et al., 2003). The result is depicted in Figure 9. The dual scaling method displays the relationship between research organizations and research areas.

In Figure 9, "translation" appears in the lower left quadrant, "learning" appears in the lower right quadrant, "statistics" and "retrieval" appear in the upper right quadrant, and "noun" and "sentence" appear in the upper left quadrant. In the vicinity around these words, the research areas and organizations relating to them appear. For example, in the upper right quadrant, Hitachi and University of Tokushima appear near "statistics" and "retrieval", which were frequent study topics for them. Similarly, "summarization" appears in the near upper right area of the source origin and is surrounded by JAIST, Toyohashi University of Technology, and Tokyo Institute of Technology., indicating it was a frequent topic of study at those institutions. We can easily see which research topics were primarily studied by each organization using this figure.

Also in Figure 9, research areas on numericals such as "probability" and "learning" appear

on the right side. Therefore, we can interpret the figure as depicting quantitative research topics on the right side and qualitative research topics on the left side. Research areas using complicated processing such as “learning” and “translation” appear in the lower area and research areas dealing with theory such as “probability”, “grammar”, “sentence”, and “noun” appear in the upper area. Therefore, we can interpret the figure as depicting theoretical research topics in the upper area and research topics using complicated processing in the lower area.

7 Conclusion

In this paper, we described a trend survey carried out on Japanese natural language processing studies done over the last ten years. We were able to investigate trend surveys on research areas very easily by treating divided words in titles by a morphological analyzer as the indications of research areas. We displayed the changes in the number of papers put out by each research organization and written on specific research topics. We also displayed the relationship between research organizations and research areas using the dual scaling method. The simple methods we used that are described here made it possible to show many useful results.

This paper has the following two significant effects:

- This paper explained a trend survey on Japanese natural language processing. By reading it, we can understand the trends in research on Japanese natural language processing. For example, we can find out which research areas were studied more often and we can see which research organizations were involved in studying natural language processing. We can also see which research organization studied a particular research area most often over the ten-year period.

- We used natural language processing to carry out the trend survey described here. For example, we automatically detected the indication of a research area from words used in titles by using a morphological analyzer. In addition, we displayed words that were extracted by the morphological analyzer in several ways to display the results of the trend survey effectively. The methods used in this paper would be useful in other trend surveys.

In short, this paper is useful for recognizing trends in Japanese NLP and for constructing methods of supporting trend surveys using NLP.

In the future, we would like to perform an international trend survey on natural language processing using international conference and journal papers such as IJCNLP, ACL, and the Journal of Computational Linguistics. We would also like to do trend surveys on other topics such as AI, biology, politics, and sociology.

The kinds of investigations we did can easily be altered to do many other kinds of investigations as well. For example, we can use the dual scaling method by investigating the relationship between the reference years and the research organizations/areas. We can also use the representation in contour for the relationship between research organizations and research areas. Although we showed the data in ascending order of the average value of the published years, we could show the data in different order, for example, the order of the total number of papers or the order of the location, i.e., showing similar research organizations/areas that are located near each other by clustering research organizations/areas using their cooccurrent words. We would like to continue to study these kinds of support methods for trend surveys in the future.

References

The Association for Natural Language Processing. 1994-2003. *Journal of Natural Language Processing*.

The Association for Natural Language Processing. 1995-2004. *Proceedings of the Annual Meeting of The Association for Natural Language Processing*.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, and Masayuki Asahara. 1999. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition.

Taichiro Ueda, Masao Karita, and Kazue Honda. 2003. *Jissen Workshop Excel Tettei Katsuyou Tahenryou Kaiseki*. Shuuwa System. (in Japanese).

Susan C. Weller and A. Kimball Romney. 1990. *Metric Scaling : Correspondence Analysis (Quantitative Applications in the Social Sciences)*. SAGE Publications.