TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
MECHANICS | MECHANIKA
2-M/2015

MICHAL PTASZYNSKI*, PAWEL LEMPA**, FUMITO MASUI*

A MODULAR SYSTEM FOR SUPPORT OF EXPERIMENTS IN TEXT CLASSIFICATION

MODULARNY SYSTEM WSPOMAGANIA DOŚWIADCZEŃ Z DZIEDZINY KLASYFIKACJI TEKSTU

Abstract

This paper presents a modular system for the support of experiments and research in text classification. Usually the research process requires two general kinds of abilities. Firstly, to laboriously analyse the provided data, perform experiments and from the experiment results create materials for preparing a scientific paper such as tables or graphs. The second kind of task includes, for example, providing a creative discussion of the results. To help researchers and allow them to focus more on creative tasks, we provide a system which helps performing the laborious part of research. The system prepares datasets for experiments, automatically performs the experiments and from the results calculates the scores of Precision, Recall, F-score, Accuracy, Specificity and phi-coefficient. It also creates tables in the LaTex format containing all the results and it draws graphs depicting and informatively comparing each group of results.

Keywords: *Experiment support, Pattern extraction, Graph generation*

Streszczenie

W niniejszym artykule przedstawiono modularny system wspomagania eksperymentów i badań z dziedziny klasyfikacji tekstu. Zazwyczaj proces badawczy wymaga dwóch typów umiejętności. Do pierwszego typu można zaliczyć wykonanie mozolnej analizy dostępnych danych, przeprowadzenie eksperymentu, a z wyników eksperymentu – stworzenie materiałów do umieszczenia w pracy naukowej, takich jak tabele lub wykresy. Drugi typ umiejętności obejmuje na przykład przeprowadzenie twórczej dyskusji na temat otrzymanych wyników. Aby pomóc naukowcom skupić się na zadaniach twórczych, w niniejszej pracy prezentujemy system, który ułatwia wykonywanie żmudnej części badań. System przygotowuje zestawy danych do eksperymentów, automatycznie przeprowadza eksperymenty, a z wyników oblicza wyniki w formie Precyzji, Zwrotu, F-miary, Dokładności, Swoistości i współczynnika phi. System tworzy również tabele w formacie LaTeX zawierające wyniki eksperymentów i rysuje wykresy porównujące każdą grupę wyników.

Słowa kluczowe: *Wspomaganie eksperymentów, ekstrakcja wzorców, rysowanie grafów*

[*] Ph.D. Michal Ptaszynski, Ph.D. Fumito Masui, Department of Computer Science, Kitami Institute of Technology.

[**] M.Sc. Pawel Lempa, Institute of Applied Informatics, Faculty of Mechanical Engineering, Cracow University of Technology.

1. Introduction

It is often said ironically about economists: “If you are so smart, why aren’t you rich?” A similar remark can be said about researchers involved in natural language processing (NLP), or computational linguistics (CL): “If you have so many language analysis and generation techniques, why don’t you use them to perform the research for you and generate a paper and presentation slides in the end?” Unfortunately, there has been astonishingly little research on scientific paper generation, presentation slides generation or even on support of the research process. One of the reasons for this is the fact that many stages of the research process require creativity, for which effective computational models still do not exist. Parts of the research which require such creative skills include for example, preparing descriptions of research background, literature review, and especially, discussion and detailed analysis of the results of experiments.

However, apart from these creative parts of research, a wide range of activities involved in the process is of a different, non-creative nature. Preparing data for experiments, conducting the experiments, step-by-step manual changing of feature sets to train and test machine learning classifiers are only some of the examples. Moreover, a thorough calculation of final scores of the evaluated tools, generating tables for the description of experiment results in technical reports and scientific papers, generating graphs from those results, and finally, description and analysis of the results – all those tasks do not require creative thinking. On the contrary, they are the non-creative part of everyday research drill. However, despite being non-creative, such activities are laborious since they require most of the researcher’s focus and precision. This could influence the motivation towards research and in practice consumes time, which could be used more efficiently for creative tasks, such as writing a detailed and convincing discussion of the results.

To help the researchers perform their research in a more convenient and efficient way we developed a system for the support of research activities and writing technical reports and scientific papers. The system is released as an Open Source set of libraries. After being initialized by one short command, the whole process including preparation of data for the experiment, conducting the experiment and generating materials helpful in writing a scientific paper is conducted automatically.

The paper’s outline is as follows. Firstly, we describe the background of our research in which we introduce a number of research studies similar to ours. Secondly, we describe the whole system. We present in detail each of the parts responsible for data preparation, experiment conduction and generation of supporting materials. Next, we present the evaluation process, which verifies the practical usability of the system. Finally, we conclude the paper and propose other features we plan to implement in the near future.

2. Related Research

The research on supporting the process of research itself is rare. The authors found only a few pieces of literature that could be considered as related to the presented system.

One of the most usable and helpful environments developed so far is the WEKA environment[1]. WEKA provides a wide range of machine learning algorithms useful in data mining tasks. It can be used as a stand-alone software, or can be called from a custom Java code to analyse data on the fly. WEKA allows data pre-processing, classification or clustering. It also provides simple visualizations of results. WEKA is widely used in the research society, especially in natural language processing (NLP) and computational linguistics (CL) fields. Unfortunately, WEKA needs especially prepared files with measurements in appropriate columns and cannot deal with plain unprocessed data (unprocessed collections of sentences, etc.). It also does not provide graphs in the format easily applicable in a research paper, nor does it provide natural language descriptions of the analysis of results.

Another tool with well-established renown is the Natural Language Toolkit (NLTK)[2]. NLTK is a Python-based platform allowing various experiments with human language data. It provides a number of tools for text classification, tokenization, stemming, and tagging, parsing, and semantic reasoning. A worth mentioning feature of NLTK is that it provides not only tools for text processing, but also a number of natural language corpora and lexical resources. A disadvantage of NLTK is that all the tools need to be launched separately. This requires at least minimal knowledge of programming languages, preferably Python, in which the toolkit was created. Thus NLTK is a very powerful and useful tool for researchers with at least minimal knowledge of programming. Another disadvantage is that not all tools included in NLTK are compatible with languages of non-alphabetic transcription (Japanese, Chinese, Korean, etc.).

In a different kind of research, Nanba et al. [1] focus on automatic generation of literature review. They assumed that in research papers researchers include short passages which can be considered as a kind of summary describing the essence of a paper and the differences between the current and previous research. Their research was very promising, as Nanba et al. [1] dealt with the creative part of research. Unfortunately, after the initial paper which presents interesting preliminary results, the method has not been developed any further. This could suggest that the creative part of the research they attempted to support, namely description of background and previous research, could still be too difficult to perform fully automatically.

Shibata and Kurohashi [2] focused on a different task, namely, on automatically generating summary slides from texts. This is not exactly the same task as creating presentation slides from a scientific paper, which we consider as one of our future tasks. However, the method they proposed, after several modifications, could be applied in our research as well. They generated slides by itemizing topic and non-topic parts extracted from syntactically analysed text. In our method the parts created by the system are grouped automatically, which could help in the itemization process.

Apart from the research described above, an interesting, although not quite scientific experiment was done by anonymous researchers involved in a campaign against dubious conferences[3]. In their attempt they generated scientific papers by picking random parts of actual papers and submitted those fake-papers to specific conferences to verify the review process of those questionable conferences. They succeeded in their task and were accepted to the conferences, which in general proved that the process of review of some conferences

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.nltk.org/

[3] https://sites.google.com/site/dumpconf/

is not of the highest quality. Therefore, if there is a similar attempt in the future, although desirably more ambitious (non-random scientific paper generation), it should submit the artificially created papers to conferences or journals of proved and well known reputation.

3. System Description

SPASS, or **S**cientific **Pa**per Writing **S**upport **S**ystem performs three tasks. Firstly, it prepares the data for the experiment, secondly, it conducts the experiment under the conditions selected by the user, and thirdly, it summarizes the results and prepares the materials for a technical report or a scientific paper. We describe each part of the process in the sections below. The general overview of the system is represented in Figure 1.

3.1. User Input

The system performs the laborious and non-creative tasks from the process of research automatically “with one click”. It is developed to help in text classification and analysis tasks. At present the system handles up to two datasets (binary classification), preferably of opposite features, such as “positive” and “negative” in sentiment analysis (SA), although the applicability of the system is not limited to SA. The user needs to prepare two separate files containing the sentences from the two corpora to be compared. The contents of these files are contrasted with each other in the process of automatic evaluation. If the input consists of only one corpus the system will simply produce the most frequent patterns (In this paper we use the words “pattern” and “n-gram” interchangeably) for the corpus.

Dataset Pre-processing. The provided sentences can be in an unprocessed form. In such a situation processed elements will consist of words (sentence tokens). However, SPASS allows any pre-processing of the sentence contents, thus making possible any kind of generalization the user might wish to apply. The experiments can be repeated with different kinds of pre-processing to check how the pre-processing influences the results. The examples of pre-processing are represented in Table 1.

Table 1

Three examples of pre-processing of a sentence in Japanese; N = noun, TOP = topic marker, ADV = adverbial particle, ADJ = adjective, COP = copula, INT = interjection, EXCL = exclamative mark

Sentence:	今日はなんて気持ちいい日なんだ！
Transliteration:	*Kyōwanantekimochiiihinanda!*
Meaning:	Today TOP what pleasant day COP EXCL
Translation:	What a pleasant day it is today!
	Preprocessing examples
1. Words:	*Kyō wa nante kimochi ii hi nanda !*
2. POS:	N TOP ADV N ADJ N COP EXCL
3. Words+POS:	*Kyō*[N] *wa*[TOP] *nante*[ADV] *kimochi*[N] *ii*[ADJ] *hi*[N] *nanda*[COP] *!*[EXCL]

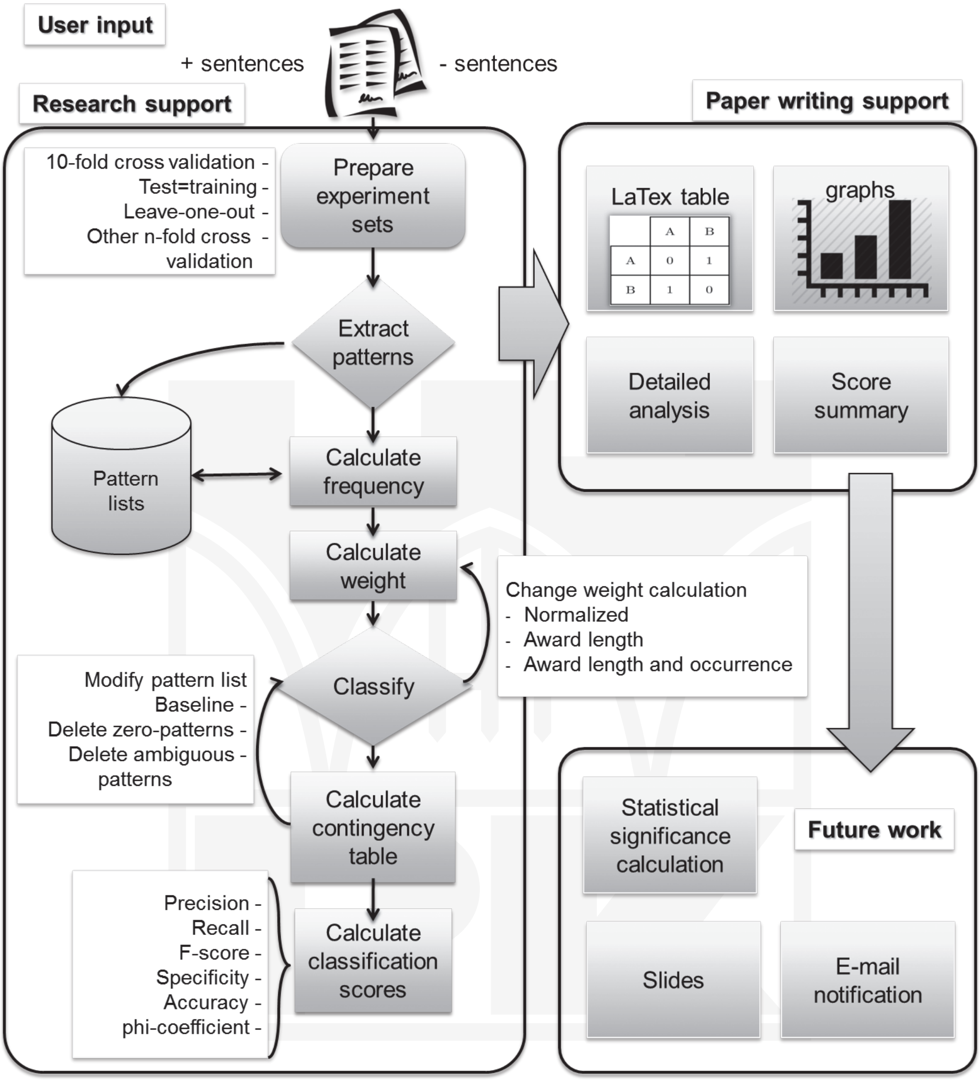


Fig. 1. General overview of SPASS divided into research support and paper writing support parts

In those examples a sentence in Japanese is pre-processed in the three following ways:

- **Tokenization**: All words, punctuation marks, etc. are separated by spaces;
- **Parts of speech (POS)**: Words are replaced with their representative parts of speech;
- **Tokens with POS**: Both words and POS information is included in one element.

In theory, the more generalized a sentence is, the less unique patterns (n-grams) it will produce, but the produced patterns will be more frequent. This can be explained by comparing tokenized sentence with its POS representation. For example, in the sentence from Table 1 we can see that a simple phrase *kimochi ii* ("feeling good/pleasant") can be represented by a POS pattern `N ADJ`. We can easily assume that there will be more `N ADJ` patterns than *kimochi ii*, since many word combinations can be represented by this morphological pattern. In other terms, there are more words in the dictionary than POS labels. Therefore, POS patterns will come in less variety but with a higher occurrence frequency. By comparing the result of classification using different pre-processing methods we can find out whether it is more effective to represent sentences as more generalized or as more specific.

3.2. Experiment Preparation Module

The initial phase in the system consists of preparation of data for the experiments. In this phase the datasets are prepared for an n-fold cross-validation test. This setting assumes that the provided data sets are first divided into n parts. Next, n-1 parts are used for training and the remaining one for testing. This procedure is performed n times so every part is used in both training and testing. The number of folds in n-fold cross-validation can be selected by the user with one simple parameter. For example, assuming the system is launched as

```
$ bash main.sh
```

The user can perform a 5-fold cross validation by adding a parameter 5, like below.

```
$ bash main.sh 5
```

The default experiment setup is 10-fold cross-validation. Setting the parameter to 1 will perform a test in which test data is the same as training data. A special additional parameter is `-loo` in which the test is performed under the “leave-one-out” (LOO) condition. In this setting all instances except one are used for training. The one left is used as a test data. The test is performed as many times as the number of all instances in the data set. For example, LOO cross validation test on a set of 35 sentences will perform the test 35 times. To speed up the process of validation, all tests are performed in parallel.

3.3. Pattern List Generation Module

The next step consists of generation of all patterns from both provided corpora. It is possible to extract patterns of all lengths. However, an informal maximum length of n-grams used in the literature is either 5-grams (applied in Google N-gram Corpus English version[4]), or 7-grams (applied in Google N-gram Corpus Japanese version[5]). This length limit was set experimentally with an assumption that longer n-grams do not yield sufficiently high frequencies. The difference between English and Japanese comes from the fact that Japanese sentences contain more grammatical particles, which means that extracting an n-gram of the same length for both languages will come with less amount of meaning for Japanese. In our system we set the default as 6-grams, although the setting can be modified freely by the users.

[4] http://catalog.ldc.upenn.edu/LDC2006T13

[5] http://googlejapan.blogspot.jp/2007/11/n-gram.html

Based on the above assumptions the system automatically extracts frequent sentence patterns distinguishable for a corpus. Firstly, all possible n-grams are generated from all elements of a sentence. From all the generated patterns only those which appear in each corpus more than once are retained as frequent patterns appearing in a given corpus. Those appearing only once are considered as not useful and rejected as pseudo-patterns. The occurrences of patterns O are used to calculate pattern weight w_j. The normalized weight w_j is calculated, according to equation 1, as a ratio of all occurrences from one corpus O_{pos} to the sum of all occurrences in both corpora $O_{pos}+O_{neg}$. The weight of each pattern is also normalized to fit in range from +1 (representing purely positive patterns) to -1 (representing purely negative patterns). The normalization is achieved by subtracting 0.5 from the initial score and multiplying this intermediate product by 2.

$$w_j = \left(\frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5 \right) * 2 \tag{1}$$

The weight is further modified in several ways. Two features are important in weight calculation. A pattern is more representative for a corpus when, firstly, the longer the pattern is (length k), and the more often it appears in the corpus (occurrence O). Thus, the weight can be modified by

- awarding length,
- awarding length and occurrence.

The formulas for modified pattern weight are represented for the "length awarded" weight w_l modification in equation 2, and for the "length and occurrence awarded" weight w_{lo} modification in equation 3

$$w_l = w_j * k \tag{2}$$

$$w_{lo} = w_j * k * O \tag{3}$$

The list of frequent patterns created in the process of pattern generation and extraction can be also further modified. When two collections of sentences of opposite features (such as "positive vs. negative") are compared, a generated list of patterns will contain patterns that appear uniquely in only one of the sides (e.g. uniquely positive patterns and uniquely negative patterns) or in both (ambiguous patterns). Therefore the pattern list can be further modified by erasing:

- all ambiguous patterns,
- only those ambiguous patterns which appear in the same number on both sides[6].

All of the above situations represent separate conditions automatically verified in the process of evaluation in the text classification task using the generated pattern lists. With these settings there is over a dozen of conditions for each of which the n-fold cross validation test is performed.

[6] Further called "zero patterns" as their normalized weight is equal to 0.

3.4. Text Classification Module

In the text classification experiment each analysed item (a sentence) is given a score. The score is calculated using the pattern list generated in the Experiment Preparation Module. There is a wide variety of algorithms applicable in text classification with which the calculation of scores can be performed. However, in the initial version of SPASS we used simple settings, which will be upgraded along the development of the system. Specifically, the score of a sentence is calculated as a sum of weights of all patterns matched for a certain sentence, like in equation 4.

$$\text{score} = \sum w_j ,\ \left(1 \geq w_j \geq -1\right) \tag{4}$$

In the future we will increase the number of applied classification algorithms, including all of the standard algorithms such as Neural Networks and Support Vector Machines. The use of a simple algorithm in the initial version allowed us to thoroughly test other parts of the system.

The score, calculated for each sentence, is automatically evaluated using sliding of the threshold window. For example, under the condition that above threshold 0 all sentences are considered positive, a sentence which got a score of 0.5 will be classified as positive. However, if the initial collection of sentences was biased toward one of the sides (e.g., more sentences of one kind, or the sentences were longer, etc.), there will be more patterns of a certain sort. Thus, to avoid bias in the results, instead of applying a rule of thumb, threshold is automatically optimized and all settings are automatically verified to choose the best model.

3.5. Contingency Table Generation Module

After the scores are calculated for all sentences the system calculates the contingency table. Depending on whether the sentence actually was positive or negative (all test sentences represent Gold Standard) the score becomes either True Positive (TP), False Positive (FP), True Negative (TN) or False Negative (FN). By calculating this for all sentences we get the contingency table for one test set (one threshold). The calculation is performed automatically for all thresholds, by sliding the threshold window by 0.1. From the contingency tables we calculate final scores using five measures. These are Precision, Recall, balanced F-score, Accuracy, Specificity and phi-coefficient. Finally, the scores are averaged for all folds from the n-fold cross validation. The average scores are a basis for further post processing.

3.6. LaTex Table Generation Module

From all the scores we generate a table in the LaTex format using a custom Perl script. The table is provided in a form already usable in a scientific paper. It contains all scores for all five measures within the whole threshold span, for experiments performed under all possible conditions (pattern list modifications and weight calculations). The table containing all information is a product of one whole single experiment and usually covers one page of an A4 or Letter type document in the LaTex format. An example of such table is represented in an Appendix at the end of the paper, and it represents the results obtained in one of the experiments in which the described system was used.

3.7. Graph Generation Module

All scores are stored in `.dat` files readable by Gnuplot[7], a standard tool for generation of high quality graphs available under most operating systems. We applied a custom Perl script to automatically generate graphs in Gnuplot for visualization of comparisons of different groups of results. One graph is generated for one kind of measure (Precision, Recall, F-score, etc.) for one compared group of results. Below we explain the compared groups of results. The graph for comparison of different weight calculations (normalized weight, length awarding, length and occurrence awarding) is drawn for:

– basic settings (with the following results compared in one graph: all patterns, length awarded all patterns, length and occurrence awarded all patterns),
– zero deleted (zero deleted, length awarded zero deleted),
– ambiguous deleted (ambiguous deleted, length awarded ambiguous deleted).

The graph for comparison of different pattern modification lists (all patterns, zero-patterns deleted, ambiguous patterns deleted) is drawn for:

– basic settings (all patterns, zero deleted, ambiguous deleted),
– length awarded (length awarded all patterns, length awarded zero deleted, length awarded ambiguous deleted).

Additionally, for all the above conditions separately, the script draws graphs containing both Precision and Recall together in one graph. This allows comparison of Break-Even Points (BEP) for all results. Also, a graph containing all results together for one measure is drawn to compare the results in a wider context. Examples of such graphs are represented in an Appendix at the end of the paper, which represents the results obtained in one of the experiments in which SPASS was used.

3.8. Result Analysis and Sentence Template Generation Module

The calculated scores are automatically analysed according to simple instructions. This module looks at the scores and compares them across the whole threshold span. It verifies the following items:

– Which modification of the algorithm was better for most of the threshold span (for all five scores);
– Which version obtained the highest BEP (in case of more than one BEP the highest one is used; calculated for Precision and Recall);
– Which version achieved the highest possible score (for all scores);
– Which version was more balanced (an algorithm which achieved high score only for one threshold is considered as generally worse than an algorithm which achieved slightly worse scores, but generally high on the whole threshold span).

[7] http://www.gnuplot.info/

Next, the results of this verification are imported into simple sentence templates, such as

> “The highest [Precision / Recall / F-score] of all was achieved by the [zero deleted / ambiguous deleted / ...] version of the algorithm”,
>
> or
>
> “When it comes to [weight calculations / pattern list modifications / ...], the highest [BEP / balanced F-score / Accuracy / ...] was achieved by [zero deleted / ambiguous deleted / ...]”.

3.9. Most Useful Pattern Extraction Module

One of the disadvantages of using standard classification algorithms, such as SVM, or Neural Networks, and making them inapplicable in traditional linguistic and corpus linguistic studies, is the fact that the analysed sentences are converted into a set of vectors. This hinders detailed analysis of the linguistic data. Therefore, we added a module allowing extraction of the most useful patterns for further linguistic analysis. During each fold in each cross validation experiment most useful patterns are extracted in the pattern matching and sentence scoring procedure. All patterns from all experiments are collected together and the patterns which appeared more than once are retained. This provides a general filtered list of patterns which where most useful during all experiments. This function has already proved to be useful by Ptaszynski et al. [3]. In their research on extracting emotive patterns from sentences they showed that patterns included in such list contained many items from their previously hand-crafted lexicon of exclamations and interjections. This suggests that it could be possible to automatically bootstrap generation of lexicons with this module. Moreover, [6] showed that such most often used pattern lists reveal not only known but also new linguistic knowledge. In their work on statistical analysis of conversations, they compared most frequent patterns from different groups of conversations between people of different age, gender, social distance and status. They found out that apart from known expressions typical for male or female interlocutors, the lists contained previously unknown patterns revealing social distance. Such patterns were not bound by any previously known linguistic rules, but in practice were used only by one group in specific conditions (e.g., only in conversations between friends, or only between people who first met).

4. Evaluation

Due to the lack of literature on research closely related to ours, no standards have been previously proposed regarding the means of evaluation of systems like the one described here. Therefore, we needed to propose our own evaluation means. One of the popular evaluation means usable in many research is a usability questionnaire. Since the system is launched by one command in a command line and the whole computation process takes place in the background, we could not ask our users questions about features such as “usability” (whether the system is easy to use), or GUI intelligibility. Instead we asked about other features or functions that might be useful to implement in the system. In the second means of evaluation we were guided by similar words to the ones opening the Introduction to this paper. Namely, “If the system is so helpful, what would be the acceptance rate for papers written with the use of it?” We understand that it is not the most objective way to evaluate the system due to many factors being involved, such as writing skills of the authors, acceptance rate of the conference etc. However, we decided to apply it as this is the most practical and an ultimate way of evaluation.

As for the first evaluation means, we performed free conversations with the present users of the system (twelve people, students and researchers of different age and career stage). From those conversations we extracted the following remarks. Having so many experiment results produced by the system it is laborious to perform statistical significance tests. One of the reasons we did not implement this feature in the initial version of the system is that there is a large number of different significance tests, depending on the type of data applied in the research. Therefore, in the future we plan to either allow users to choose their test, or, which would be more desirable, find a method for automatic selection of a statistical significance test depending on the data. Another useful function would be to generate presentation slides, at least partially. This could be easily implemented as the system generates all materials in a LaTex template. The third function worth implementation would be email notification when the whole process is finished. Depending on the amount of data, the whole process could take a few seconds, an hour, or even a day or more, especially, when the user tries to analyse BigData. It could be tiresome to sit and wait for the results. Therefore an email notification would be a useful feature. However, this would mean the necessity of additional settings (ensuring the server has appropriate generic software to send a simple email message), while initially we meant the system to work “out of the box”.

As for the second means of evaluation, at the time of writing, there have been several papers accepted to different conferences written with the use of SPASS. First three of them analyse emotional and non-emotional sentences [3–5]. In this research our system helped confirm that completely automatic approach to extraction of emotional patterns from sentences can give similarly good results to tools developed manually. In the second set of publications [6, 7], the system was applied in a conversation analysis task to find similarities in conversations between interlocutors of different age, gender, social distance and status. Interestingly, the system extracted several linguistic rules (confirmed statistically) which were previously unknown. Finally, in the last set of publications [8, 9] the system was applied in the analysis of future related expressions for the task of future prediction from trend information. The experiments performed by the system helped prove that sentences referring to the future contain frequent patterns, while patterns in other sentences (non-future related, such as present, past or not time related) are sparse and scattered. This proved that “future-referring sentences” can be treated and analysed as one separate kind of sentences. This discovery helped Nakajima et. al [8] choose appropriate methods for further analysis of their data (e.g., grounded in linguistics rather than in information extraction, or data mining).

5. Conclusions

Research is a process requiring two kinds of abilities – creativity and precision. The tasks requiring creativity include preparing detailed analysis of experiment results or writing a convincing discussion. The tasks which are not creative, but requiring focus and precision include laborious preparation of data for experiments, performing the experiments and preparing materials for writing a scientific paper, such as tables or graphs. Computers are poor at creative tasks, but good at laborious non-creative tasks. People on the other hand are experts when it comes to creativity, but the passion for research could be severely impaired by the laborious tasks included in the everyday research drill. Therefore to ease the researchers, and allow them focus on the creative part of the research process, we developed SPASS – a system which helps performing the laborious part of the research. SPASS is a system for the support of research and writing of scientific papers. The system prepares the data for the experiments, automatically performs the experiments and from the results calculates the scores according to five different kinds of measures (Precision, Recall, etc.). It also creates tables in the LaTex format containing all the results, draws graphs depicting and informatively comparing each groups of results and generates descriptions of those results using sentence templates. And what is most important, it does all that with one single command.

6. Future work

In the near future we plan to upgrade the system further and implement additional functions. First of all, we plan to add various classification algorithms for more thorough evaluation. We also plan to include automatic calculation of statistical significance of results. We also plan to perform the n-fold cross validation multiple times to further improve the objectivity of the results. A useful function would be an e-mail notification about the finalization of the whole experiment process so the researchers did not have to wait for the results. When it comes to the descriptions of experiment results, at this point they are generated as generic sentences containing certain knowledge. In the near future we plan to perform automatic summarization of sentence templates to increase the readability and informativeness of the descriptions. This would move the research from paper writing support one step toward an actual automatic paper generation. We also plan to implement a script for generation of presentation slides in the LaTex template from the results description, similarly to Shitaba and Kurohashi [2]. At present the system handles only two classes of labels, or two corpora differing in a certain feature (positive/negative). In the future we plan to perform different kinds of data as well, especially multi-label classification, with classes either related to each other, or unrelated. This could help deal with not only binary classification-like corpora comparison (such as sentiment analysis), but also wider scale analysis, such as extracting expressions specific for certain emotion types (fear, sadness, joy etc.), or gradual sentiment (for example product reviews on Amazon[8]). This would help extract pragmatic generalizations from corpora, similarly to Potts and Schwarz [10], and could contribute greatly to the emerging field of computational pragmatics.

[8] http://www.amazon.com/

References

[1] Hidetsugu Nanba, Noriko Kando, Manabu Okumura, *Classification of research papers using citation links and citation types: Towards automatic review article generation*, [in:] Proceedings of 11th ASIS SIG/CR Classification Research Workshop, 2000, 117-134.

[2] Tomohide Shibata, Sadao Kurohashi, *Automatic slide generation based on discourse structure analysis*, [in:] Proceedings of IJCNLP, Springer Berlin Heidelberg, 2005, 754-766.

[3] Ptaszynski M., Fumito Masui, Rzepka R., Kenji Araki, *Automatic Extraction of Emotive and Non-emotive Sentence Patterns*, [In:] Proceedings of The Twentieth Annual Meeting of The Association for Natural Language Processing (NLP2014), Sapporo, Japan, 2014, 868-871.

[4] Ptaszynski M., Fumito Masui, Rzepka R., Kenji Araki, *Emotive or Non-emotive: That is The Question*, [in:] Proceedings of 5th Workhsop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2014), Baltimore, USA, 2014, 59-65, held in conjunction with The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), June 22–27.

[5] Ptaszynski M., Fumito Masui, Rzepka R., Kenji Araki, *Detecting emotive sentences with pattern-based language modelling*, [in:] Proceedings of the 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems – KES2014, 484-493, Gdynia, Poland 2014.

[6] Ptaszynski M., Dai Hasegawa, Fumito Masui, Hiroshi Sakuta, Eijiro Adachi, *How Differently Do We Talk? A Study of Sentence Patterns in Groups of Different Age, Gender and Social Status*, [in:] Proceedings of The Twentieth Annual Meeting of The Association for Natural Language Processing (NLP2014), Sapporo, Japan 2014, 3-6.

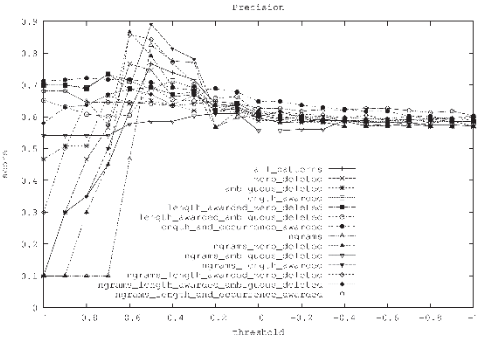
[7] Ptaszynski M., Dai Hasegawa, Fumito Masui, *Women Like Backchannel, But Men Finish Earlier: Pattern Based Language Modeling of Conversations Reveals Gender and Social Distance Differences*, [in:] Samsung HLT Young Researchers Symposium, 9th International Conference on Natural Language Processing (PolTAL 2014), paper ID: 02, Warszawa, Poland 2014.

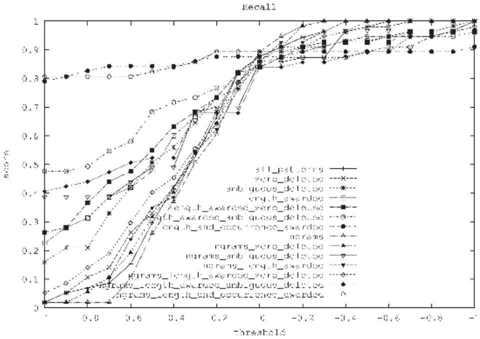
[8] Yoko Nakajima, Ptaszynski M., Hirotoshi Honma, Fumito Masui, *Investigation of Future Reference Expressions in Trend Information*, [in:] Proceedings of the 2014 AAAI Spring Symposium Series, "Big data becomes personal: knowledge into meaning For better health, wellness and well-being", Stanford, USA 2014, 31-38.

[9] Yoko Nakajima, Ptaszynski M., Hirotoshi Honma, Fumito Masui, *Extraction of Future Reference Expressions in Trend Information*, [in:] The Proceedings of 24th Intelligent System Symposium (FAN2014), Kitami, Japan 2014.

[10] Potts Ch., Schwarz F., *Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora*, Ms., UMass Amherst, 2008.

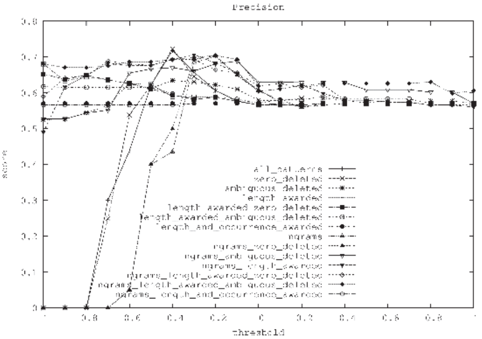
Appendix 1: Graphs representing experiment results from Ptaszytnski et al. [4]

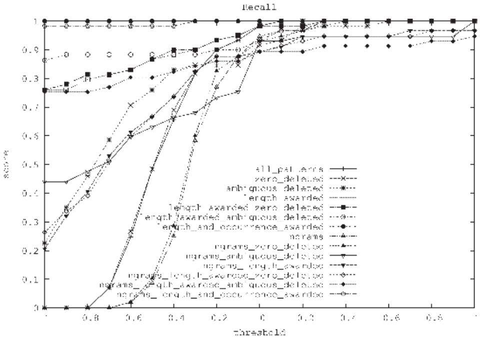




(a) Precision comparison for tokenized dataset.

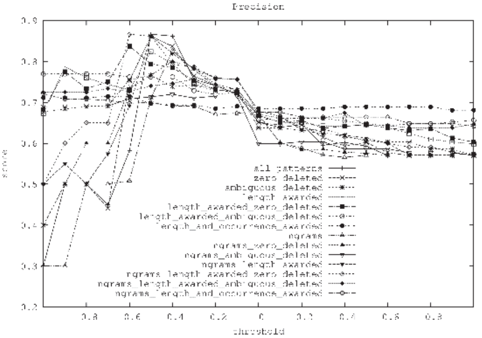
(b) Recall comparison for tokenized dataset.

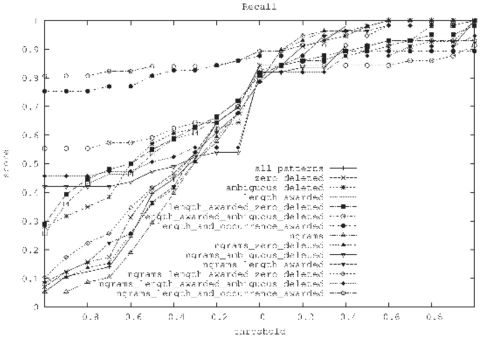




(c) Precision comparison for POS-tagged dataset.

(d) Recall comparison for POS-tagged dataset.





(c) Precision comparison for tokenized dataset with POS tags.

(f) Recall comparison for tokenized dataset with POS tags.

Appendix 2: Experiment results from Ptaszynski et al. [4], corresponding to graphs in Appendinx 1

ngrams

Threshold	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00	-0.10	-0.20	-0.30	-0.40	-0.50	-0.60	-0.70	-0.80	-0.90	-1.00
Precision	0.10	0.10	0.10	0.10	0.47	0.82	0.77	0.77	0.57	0.60	0.60	0.59	0.59	0.58	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Recall	0.02	0.02	0.02	0.02	0.10	0.26	0.37	0.51	0.61	0.78	0.89	0.95	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F-score	0.03	0.03	0.03	0.03	0.17	0.40	0.50	0.61	0.59	0.68	0.72	0.73	0.73	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.73

ngrams zero deleted

Threshold	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00	-0.10	-0.20	-0.30	-0.40	-0.50	-0.60	-0.70	-0.80	-0.90	-1.00
Precision	0.10	0.10	0.30	0.45	0.87	0.79	0.70	0.70	0.57	0.62	0.60	0.58	0.59	0.58	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Recall	0.02	0.02	0.06	0.11	0.19	0.32	0.40	0.54	0.64	0.82	0.88	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F-score	0.03	0.03	0.10	0.17	0.32	0.45	0.51	0.61	0.60	0.71	0.71	0.71	0.74	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.73

ngrams ambiguous deleted

Threshold	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00	-0.10	-0.20	-0.30	-0.40	-0.50	-0.60	-0.70	-0.80	-0.90	-1.00
Precision	0.54	0.54	0.54	0.54	0.58	0.59	0.59	0.60	0.61	0.61	0.56	0.56	0.56	0.56	0.59	0.58	0.58	0.58	0.57	0.57	0.57
Recall	0.39	0.39	0.39	0.39	0.44	0.49	0.49	0.68	0.70	0.70	0.86	0.86	0.87	0.87	0.97	0.97	0.97	0.97	0.97	0.97	1.00
F-score	0.45	0.45	0.45	0.45	0.50	0.53	0.53	0.64	0.65	0.65	0.67	0.67	0.68	0.68	0.73	0.73	0.73	0.73	0.72	0.72	0.73

ngrams length awarded

Threshold	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00	-0.10	-0.20	-0.30	-0.40	-0.50	-0.60	-0.70	-0.80	-0.90	-1.00
Precision	0.10	0.30	0.35	0.50	0.69	0.89	0.81	0.78	0.61	0.63	0.60	0.62	0.60	0.60	0.58	0.58	0.58	0.58	0.57	0.57	0.57
Recall	0.02	0.05	0.07	0.10	0.24	0.35	0.39	0.54	0.62	0.76	0.86	0.93	0.93	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F-score	0.03	0.09	0.12	0.17	0.36	0.50	0.52	0.64	0.62	0.69	0.71	0.74	0.73	0.74	0.74	0.74	0.74	0.73	0.73	0.73	0.73

ngrams length awarded zero deleted

Threshold	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00	-0.10	-0.20	-0.30	-0.40	-0.50	-0.60	-0.70	-0.80	-0.90	-1.00
Precision	0.30	0.50	0.65	0.57	0.86	0.84	0.77	0.70	0.62	0.62	0.61	0.62	0.61	0.61	0.60	0.61	0.59	0.58	0.57	0.57	0.57
Recall	0.05	0.09	0.14	0.19	0.30	0.40	0.45	0.56	0.64	0.79	0.86	0.91	0.91	0.95	0.96	1.00	1.00	1.00	1.00	1.00	1.00
F-score	0.09	0.15	0.23	0.29	0.44	0.55	0.57	0.62	0.63	0.70	0.71	0.73	0.73	0.74	0.74	0.75	0.74	0.74	0.73	0.73	0.73

ngrams length awarded ambiguous deleted

Threshold	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00	-0.10	-0.20	-0.30	-0.40	-0.50	-0.60	-0.70	-0.80	-0.90	-1.00
Precision	0.58	0.63	0.64	0.67	0.69	0.67	0.64	0.67	0.64	0.63	0.62	0.61	0.60	0.59	0.60	0.59	0.59	0.60	0.60	0.60	0.60
Recall	0.41	0.42	0.44	0.47	0.51	0.52	0.52	0.68	0.68	0.68	0.84	0.84	0.86	0.86	0.88	0.89	0.89	0.95	0.95	0.97	1.00
F-score	0.48	0.51	0.52	0.55	0.58	0.59	0.57	0.67	0.66	0.65	0.71	0.70	0.71	0.70	0.71	0.71	0.71	0.74	0.73	0.74	0.75