Computer Aided Writing

A Prototype of an Intelligent Word Processing System

André Klahold, Johannes Zenkert and Madjid Fathi
Institute of Knowledge Based Systems and Knowledge Management
University of Siegen
Siegen, Germany
{andre.klahold, johannes.zenkert, madjid.fathi}@uni-siegen.de

***Abstract*— There are more and more capable computer based concepts feasible for the writing process. In the coming years a new generation of tools supporting text production will originate. They will by far exceed modern word processing software and change the writing process itself fundamentally. This work presents a prototype of an intelligent word processing system "VAC" aimed at "Computer Aided Writing". It gives a glimpse of what will be and an impression on what already is.**

Keywords— Computer Aided Writing; Recommender Systems; KDT; Word Associations

I. INTRODUCTION

The roots of Computer Aided Writing are to be found in the 1950s. The first use of a computer to support journalistic work was actually intended as a PR-event. As the person responsible for CBS News, Journalist Walter Cronkit, stated:

"It was agreed that it would be used on our election night purely, quite frankly, almost as a gimmick, to try to introduce the American people to what these machines could do, and also to give them some added excitement on election night. I thought it was pretty much gimmickry. I didn't see the great potential of them despite the propaganda put out by the UNIVAC people and the others." [1]

The writing process is constituted by three main phases: Planning, Translating and Reviewing. Hayes and Flower [2] established a cognitive theory of writing. Their model has three main components:

(1) The task environment includes all relevant external conditions regarding the writing process

(2) The writers long term memory as the internal basis

(3) The writing process itself is divided in the three phases

Before we introduce the computer-based concepts to support the writing process, we will have a closer look at these three phases of the writing process.

Subsequently, we present our computer-aided writing prototype "VAC" and give an overview of its text processing capabilities and supportive functionalities in the initial planning phase of the writing process. The conclusion of this paper summarizes the results of our prototype and provides an outlook on future work.

II. PHASES OF THE WRITING PROCESS

Fig. 1: The three phases of the writing process, introduced by Hayes and Flower [2], are not strictly linear, but more like sub-processes each influencing each other all the time.

A. Planning

In the *Planning* process [2] an internal representation of an idea is build. The author uses his long term memory and archived information to develop ideas. Afterwards, he has to organize his ideas. And finally the "Goal-setting" is necessary. The author creates these goals in the same process that generates ideas.

The planning phase is also influenced by the current context of the author. And last but not least, active search for additional information is one of the cornerstones of the planning phase.

Machill et al. [3] conducted a study about research behavior in professional editorial environments. Their result is that with 47%, computers are the most used research instrument. The time spent on search engines accumulates to 4.1%. In this segment Google is nearly the only used tool (especially in Germany with 99.3%) [3]. Specific websites are used with nearly the same amount of time. Most writers use only ten websites for about 40% of their research [3].

978-1-5386-2652-8/17 $31.00 © 2017 IEEE
DOI 10.1109/CSCI.2017.37

B. Translating

As Flower and Hayes [2] state, *"Translating is the process of putting ideas into visible language"*. Ideas for example, could be mental pictures, keywords or drawings. Even if the result of the planning process is represented by words it is unlikely that they follow syntax or other rules for written text.

The process of translating needs the writer to generate syntactic, lexical and temporal correct sentences. The sentences must be in a specific order following syntactic but also semantic rules. This task may demand so much of the writer's mental capacity that the main goal of translating the ideas could suffer.

Today's word processors support the writer during the translating phase to reduce his mental load. Text processors arose with the replacement of typewriters with computers. Already in 1965 Magnuson [4] described a prototype, which supported the separation of texts and sentences in the sense of typesetting. A recommendable overview about the history of computer aided writing is given by Haigh [5].

C. Reviewing

The reviewing process consists of evaluating and revising. In this phase the writer reads what he has written. Text adaptions and corrections are made and then again revised.

III. A Computer Aided Writing Tool

Today we already use machine support in the different phases of text creation. For example, existing text processing software supports our own creativity in text production, especially in the aforementioned translating and reviewing phases. Typically, users of these comprehensive software solutions and tools are supported with a variety of functions which allow writing and formatting of text. Auto correction, spelling checking and automatic grammar checking are also common and very useful instruments in the translating phase. The available text processing solutions have a user-oriented background, should be easy to use and flexible in their application. In relation to computer-aided writing, the mental capabilities of the user remain highly stressed to actually produce meaningful text with existing tools.

As a conclusion of today's text processing applications and tools, we see that the translating phase is already well supported. But the two other phases are not so well equipped.

A. Computer Aided Writing - Prototype "VAC"

Our Computer Aided Writing prototype ("VAC") addresses the lack of pre-writing functionalities and its primary goal is to help writers during the planning phase.

VAC implements different text mining methods, offers knowledge extraction tools which facilitate the information gathering process and supports the initial text production. Nevertheless, VAC provides some new types of tools for author assistance during the translating and reviewing phases, too.

The basis of VAC is a knowledge-based system that processes, utilizes and recommends text information, especially from news or semantically related content such as Linked Open Data (LOD) sources. In the processing of information and integration of different text mining results, VAC follows a multidimensional pre-processing and knowledge representation approach to handle text mining results.

The idea of integrative text mining and handling of method result has been described in our previous work [6].

B. Knowledged Processing and Representation

VAC follows an integrative text mining approach which combines the results of different text mining methods in a multidimensional knowledge representation framework [6].

We have developed this approach to process textual content and extract meaningful information with a combination of different existing text mining methods. In this way, contained information related to (multi-level) sentiment, (multi-) topics, named entities and associative relationships are extracted and represented in the knowledge base.

We consider the representation framework advantageous compared to other entity-oriented representation frameworks, because it directly integrates potential text mining analysis results and therefore requires less re-calculation in visualization tasks or search requests.

Through various transformation operations like dimensional filtering, reduction or selection, analysis results are adapted accordingly and query-specific results are easily retrieved from the knowledge base in desired knowledge presentation or visualization formats.

IV. Support during the Planning Phase

A. Collecting and structuring Ideas

The ICP component ("idea cooking pot") of the VAC system lets the writer collect and structure his ideas. The basic function of this tool is data collecting, storing and presentation. So far it is merely a database and not intelligent in any way. But when an idea, represented by a title and a description, is inserted in the ICP, a simultaneous integrative analysis process is running. It uses KDT (Knowledge Discovery from Text) techniques to extract the following information:

- Named Entities (person, place, event, organization, et cetera)
- Temporal information (date, period, et cetera)
- Keywords (the central words of the text from a statistical point of view) and
- Semantically related documents

The extracted information is continually updated and displayed for the author while he is typing. So a lot of information the writer normally has to search for is automatically available. For example, the details of one person's vita curriculum he is writing about (see Fig. 2). Thereby information is transformed into knowledge by presenting it actively in the right context.

Fig. 2: The ICP component of the VAC systems provides automatic information regarding the ideas created by the author.

The named entity recognition [7] is carried out with the help of a pattern recognition method as well as a dictionary based on DBPedia [8]. In essence the named entity candidates found in the description of an idea are tested by the semantic intersection between the DBPedia text belonging to that specific entity and the ideas description. To calculate the semantic overlay we use the extended *CRIC* method proposed in [9].

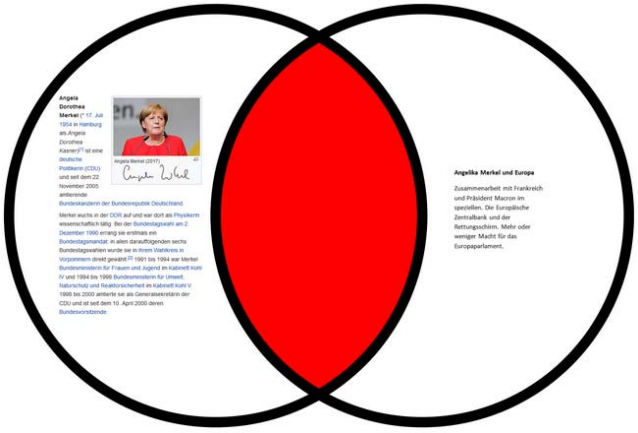

Fig. 3: Semantic intersection between DBPedia description (left) and the ideas description (right). Only significant intersections induce positive entity recognition.

Temporal information is extracted via rules. The keyword extraction and recommendation of related articles also uses the extended *CRIC* method.

Given that a user or a group of users store all their ideas in the VAC system, we use the information about the semantic proximity of each related article and calculate a "uniqueness" value for each idea.

By doing this, the ICP component is able to inform the writer if an idea is redundant. Given this information the writer is able to adapt his ideas in a very early stage of the planning phase and thereby spare him useless effort.

"A recommender system is a system, which actively offers a subset of *beneficial* elements to a user with a specific context" (translated definition based on [10]).

Therefore, the ICP component is a recommender system. When we tested the ICP component of the VAC system with a test user group this feature was the must appreciated one.

B. Surveillance of new information

Besides active recommendations the VAC systems offers a TD (Topic Discovery) component for research purposes. It is based on a multitude of configurable content sources. That could be news sites or scientific papers for example and should support the writer's topics. An indexing function continually monitors all sources and gathers new content instantly. Based on that information, a word cloud (with customization possibilities) representing the information of the day is build and displayed. Figure 4 and Fig. 5 provide examples.

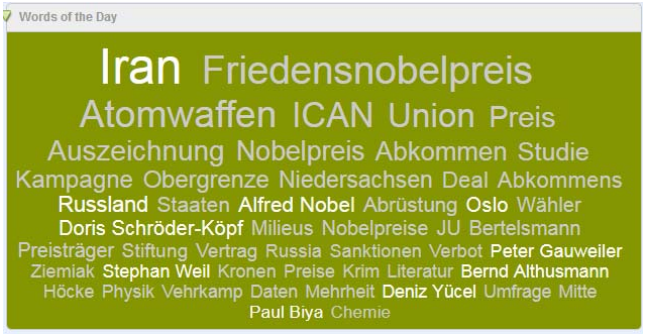

Fig. 4: The "words of the day" function of the topic discovery component builds a cloud of the most prominent words of the day.

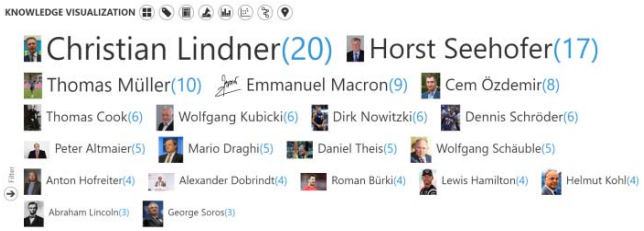

Fig. 5: Semantically related information allows further filtering, e.g. only today's information with occurences of named entities with type "person".

For all words per source their frequency per day (date of text generation) is logged. Based on this, the average value and standard deviation can be computed, which then serve as point for words with the characteristic „regular usage". We then assumed that words, whose usage frequency in relation to their

regular use (see Fig. 6) in one period rises strongly, are to be concerned "important".

After numerous experiments the second derivative of the average value of the frequency proved as the best indicator. The algorithm implemented calculates the most important word of n words of the day by building the second derivate of the average frequency for day t. We use the frequency of last five days ($t = 1, \dots, 5$) and build the sum of the differences:

$$Max_{i=1\dots n}\left(f''_t(average\_frequeny_t(word_i)\right) \quad (1)$$

That is comprehensibly, since on the one hand the average value absorbs fluctuations in frequency, on the other hand the first derivative of the frequency reflects their growth and the second derivative reflects the growth rate. Words, whose frequency rises fast in one period are particularly important.

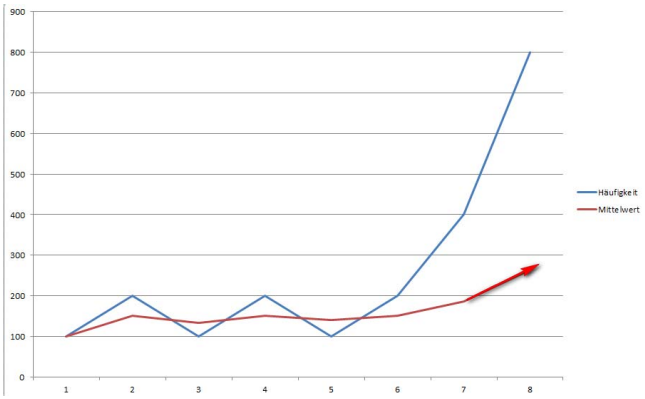

Fig. 6: Computing the important word of the day by using the second derivative of the average value of the frequency (red curve).

Also all the articles are displayed. To give the user an easier access to potential big amounts of articles they are grouped. We use the topic detection developed in [11] and semantic intersection to form the groups.

Fig. 7 shows an overview of VAC's functionality "topics of the day". In the example shown, German news articles from different news sites are grouped based on their content and detected topic.

The nobel peace prize, the analysis of results from German federal election, the Catalan independence referendum and discussions between USA and Iran have been detected as topics and are shown in Fig. 7.

As stated before, each topic can be represented by one or more related knowledge items which are grouped to give an improved overview. Each group is collapsible in VAC's interface and shows or hides a list of related articles and further information.

Fig. 7: Articles are grouped by topic for easier access

C. *Search Assistance*

If the writer uses a specific search phrase, which could be "clicked" by choosing words from the word cloud, there is further support given by the TD component. In the following, may the word "Friedensnobelpreis" (nobel peace prize) be the word which the user is interested in and which has been selected by the user in the following scenario.

First, within VAC there is a timeline displayed. A corresponding histogram of the word appearance shows the amount of related information and occurrence of the specific search phrase (or keyword) during a specific day (represented in an hour scaling). The time line is shown in Fig. 7.

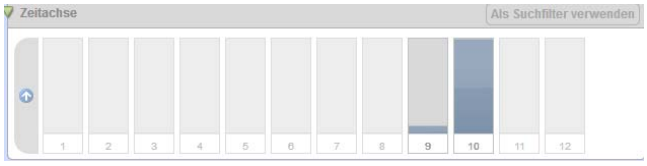

Fig. 8: Timeline for a search word

Second, word associations are derived based on a previous developed algorithm, which resembles human word associations [12]. The CIMAWA term is defined as follows

$$CIMAWA^{\zeta}_{ws}\left(x(y)\right) = \frac{Cooc_{ws}(x,y)}{(frequency\ (y))^{\alpha}} + \zeta * \frac{Cooc_{ws}(x,y)}{(frequency\ (x))^{\alpha}} \quad (2)$$

where $\mathrm{CIMAWA}^{\zeta}_{ws}(x(y))$ is a measure for indicating the strength of the word 'x' in association with the word 'y', based on a certain window-size (*ws)* for the co-occurrence *(Cooc(x,y))* measuring and a damping factor ζ. The co-

occurrence is a statistical measure that expresses how many times two words coexist in a defined text window.

We improved the concept for word association presented in [12] by a long and short term memory association. We achieved this by a daily shifting division of the corpus. The association is calculated in a "long term" and a "short term" part of the corpus. The later contains all content that is not older than 15 days. All the other content resides in the "long term" part of the corpus. Each day, additional content older than 15 days is moved into this part. While moving the content, the co-occurrence values are corrected accordingly.

As evident in Fig. 9 the calculated long term association (mid) connects the selected word "Friedensnobelpreis" (Nobel Peace Prize) with the 2016 laureate Juan Manuel Santos and his country, Columbia. Also 2014 laureate Malala Yousafzai and the city where the committee resides are associated in long term.

Short term associations (bottom) are the 2017 laureate "ICAN" and the words "atomic weapons", "Nobel Prize", "Nobel Prize committee", "campaign" and "honor".

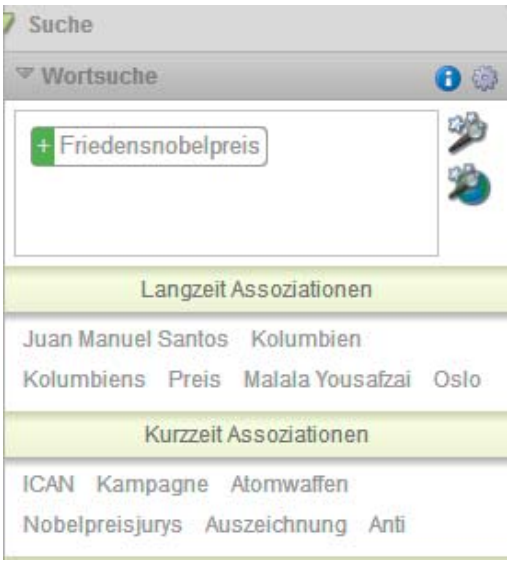

Fig. 9: An example for associations build based on the word "Friedensnobelpreis" (Nobel Peace Prize) on the 7th of October 2017.

Word associations help the writer to further specify his information needs. If the Peace Nobel Prize for laureate Mrs. Yousafzai is what the author has in mind, one click on the association refines the search and selects helpful content as shown in Fig. 10.

The utilization of word associations as context-aware associative search instrument has been also proposed in our previous work [13]. Word association strength has been used as distance measure (Associative Proximity Measure) to organize the visualization of nodes and other knowledge items in knowledge maps [13].

Fig. 10: An example for associations build based on the word "Friedensnobelpreis" (Nobel Peace Prize) on the 7th of October 2017.

V. Conclusion

This paper presents a prototype of an intelligent word processing system "VAC" aimed at "Computer Aided Writing". It supports authors especially during the Planning phase of the writing processes. VAC implements different text mining methods to pre-process textual content and extract meaningful information into the knowledge base.

The ICP and TD components are the main building blocks. ICP helps collecting and structuring ideas. The TD component provides automatic surveillance of new information as well as a search interface which helps the user to refine his search by proposed word associations.

A few similar approaches exist, but use other methods to realize a Computer Aided Writing concept. Liu et al. [14] propose a system for writing an assisted love letter with the help of keyword extraction, sentence construction and synonym substitution. In [15], Liu et al. develop this approach further to assist blog writing.

In future work towards the direction of Computer Aided Writing, we are going to further develop our VAC prototype to support all three phases of the writing process with idea creation, discovery, searching and recommendation components.

References

[1] D. P. Julyk, "The Trouble With Machines Is People. The Computer as Icon in Post-War America: 1946-1970", The University of Michigan. 2008.

[2] J. R. Hayes and L. Flower, "A Cognitive Process Theory of Writing", College Composition and Communication, 32(4), pp. 365–387, 1981.

[3] M. Machill, M. Beiler, and M. Zenker, "Journalistische Recherche im Internet. Bestandsaufnahme Journalistischer Arbeitsweisen in Zeitungen, Hörfunk, Fernsehen und Online", Schriftenreihe Medienforschung, Landesanstalt für Medien NRW 60, pp. 164–304, 2008.

[4] R. A. Magnuson, "Automated Documentation", Research Analysis Corporation McLean, Virginia, USA, pp. 1–13, 1965.

[5] T. Haigh, "Remembering the office of the future: The origins of word processing and office automation", IEEE Annals of the History of Computing 28(4), pp. 6–31, 2006.

[6] J. Zenkert and M. Fathi, "Multidimensional knowledge representation of text analytics results in knowledge bases", 2016 IEEE International Conference on Electro Information Technology (EIT), Grand Forks, ND, USA, pp. 541-546. 2016

[7] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification”, Linguisticae Investigations 30, pp. 3–26, 2007.

[8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, “DBpedia: A Nucleus for a Web of Open Data”, In: K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg, 2007.

[9] A. Klahold, „CRIC: Kontextbasierte Empfehlung unstrukturierter Texte in Echtzeitumgebungen“, Dissertation, University of Siegen, 2006.

[10] A. Klahold, „Empfehlungssysteme. Grundlagen, Konzepte und Systeme“, pp. 1–188. Vieweg + Teubner, Wiesbaden, 2009.

[11] A. Klahold, P. Uhr, F. Ansari, and M. Fathi, “Using word association to detect multitopic structures in text documents”, IEEE Intelligent Systems, 29(5), pp. 40-46, 2014.

[12] A. Klahold, P. Uhr, and M. Fathi, “Imitation of the Human Ability of Word Association as a Basis for Topic Detection”, IEEE Transactions on Knowledge and Data Engineering, Status in Review, 2012.

[13] J. Zenkert, A. Holland, and M. Fathi, “Discovering contextual knowledge with associated information in dimensional structured knowledge bases”, 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 2016.

[14] C. Liu, C. Lee, S. Yu, and C. Chen, “Computer assisted writing system”, Expert Systems with Applications 38(1), pp. 804–811, 2011.

[15] C. L. Liu, W. H. Hsaio, C. H. Lee, and H. C. Chi, “An HMM-based algorithm for content ranking and coherence-feature extraction”, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43(2), pp. 440-450, 2013.