Affect analysis in context of characters in narratives

Michal Ptaszynski, Hiroaki Dokoshi, Satoshi Oyama, Rafal Rzepka, Masahito Kurihara, Kenji Araki, Yoshio Momouchi

1. Introduction

Understanding emotions expressed in language is an important part of communication between humans. Along with the development of language and writing people were able to develop various strategies for expressing emotions through text. Although one of the most obvious situations in which we express emotions is a conversation, the emotions have been expressed in many different types of language vehicles. Some of the examples include epistolary art (e.g., love letters), personal diaries (with their more recent version – blogs) or stories (e.g., fairy tales). These all are part of one category, namely narratives. What differentiates narratives from conversation is its descriptive force. While in conversations most of the time the subject is the first person ("me/I"), narratives provide an opportunity to describe a situation through the eyes of a third person. Especially in children stories and fairy tales this becomes an apparatus to teach the readers (children) important emotional strategies. For example, the fear elicited toward Little Red Riding Hood being devoured by the Bad Wolf helps children understand that they should not talk to strangers. Or when the little Goat Child, from the story of "The Wolf and the Seven Young Kids", is saved from being eaten and runs to get help, children can understand the value of relief and happiness after a difficult struggle.

Understanding emotions in narratives, and especially in children stories, is therefore a crucial task to develop a computational model of language understanding. In this paper we address the task of recognition of emotions (later called affect analysis) in children stories. In general, affect analysis within the context of narratives is a challenging task because narratives are created of different kinds of sentences (descriptions, dialogs, etc.). Moreover, different characters become subjects of different emotional expressions in different parts of narratives. In this research we address the problem of person/character related affect recognition in narratives. We propose a method for emotion subject extraction from a sentence based on analysis of anaphoric expressions and compare two methods for affect analysis. We evaluate the system and discuss its possible future improvements.

© 2012 Elsevier Ltd. All rights reserved.
In Section 3 we describe the previously developed tools we apply in our methods. In Section 4 we present the method allowing simultaneous affect analysis of separate characters in narrative. We evaluate the method in Section 5. Finally, in Section 6, we conclude the paper and propose some ideas to improve the described method as well as its possible future applications.

## 2. Affect analysis in narratives

Text based affect analysis (AA) has been defined as a field focused on developing natural language processing techniques for estimating the emotive aspect of text (Grefenstette, Qu, Shanahan, & Evans, 2004). For example, Elliott (1992) proposed a keyword-based Affect analysis system applying an affect lexicon (including words like “happy”, or “sad”) with modifiers (“extremely”, “some-what”). Liu, Lieberman, and Selker (2003) presented a model of text-based affect sensing based on OMCS (Open-Mind Common Sense), a generic common sense database, with an application to e-mail interpretation. Alm, Roth, and Sproat (2005) proposed a machine learning method for affect analysis of fairy tales. Aman and Szpakowicz also applied machine learning techniques to analyze emotions expressed on blogs (Aman & Szpakowicz, 2007).

There have also been several attempts to achieve this goal for the Japanese language. For example, Tsuchiya, Yoshimura, Watabe, and Kawaoka (2007) tried to estimate emotive aspect of utterances with the use of an association mechanism. On the other hand, Tokuhisa, Inui, and Matsumoto (2008) as well as Shi, Rzepka, and Araki (2008) used a large number of examples gathered from the Web to estimate user emotions. Furthermore, Ptaszynski, Dybala, Shi, Rzepka, and Araki (2009a) proposed a Web-based supported affect analysis system for Japanese text-based utterances.

There have been also some research in affect analysis of narratives. For example, Alm et al. (2005) used a machine learning algorithm with multiple linguistic features to distinguish general affect expressed in stories, especially in fairy tales. Neviarouskaya, Prendinger, and Ishizuka (2010) tested their affect analysis model for English on various kinds of data, including fairy tales (the same stories as Alm et al.). Finally, Mohammad (2011) focused mostly on variations of speech signals and simplified functions to measure actor level emotion magnitude. However, their results were less than ideal due to lacks in the lexicon and inconsistencies in syntactic rules. The previous research for affect analysis of narratives described above is summarized in Table 1.

In all of the above research on affect analysis in narratives the analysis was simplified to detecting general affect for a story, or part of the story (sentence/paragraph). In this research we focused not on general affect, but tried to specify who is the subject of the particular emotion.

Previously there have been a few research on affect analysis of emotion subjects (characters/actors), both for English language. Zhang (2010) proposed affect analysis method for text-based conversations, in which she used predetermined relationship profiles. The final emotion score calculated for user utterance was influenced by the type of relationship the user had with previous speaker. For example, if user A was in a bad relationship with previous speaker B it was more likely A would express negative emotion, if B expressed positive emotion, like joy, or happiness. Another research was done by Calix and Knapp (2011). They tested linear vs. non-linear Support Vector Regression (SVR) with different kernel functions to measure actor level emotion magnitude. However, they focused mostly on variations of speech signals and simplified textual features to keyword spotting.

Comparing to all previous research, the research presented here is pioneering in several ways. Firstly, it is the first attempt to combine the tasks of text-based affect detection in fairy tales and in characters (emotion subjects). It is also the first attempt to apply thesaurus-based methods to each of these tasks and compare thesaurus-based methods (based on Japanese WordNet) to corpus-based methods (based on ML-Ask system). Finally, it is the first attempt to perform all of the above in Japanese language.

### 3. Tools

In this section we describe all tools and methods for basic affect analysis (AA) used further in contextual affect analysis tasks.

#### 3.1. Emotive expression dictionary

Emotive Expression Dictionary (Nakamura, 1993) is a dictionary developed by Akira Nakamura in a period of over 20-year time. It is a collection of over two thousand expressions describing emotional states collected manually from a wide range of literature. It is not a tool per se, but was converted into an emotive expression database by Ptaszynski, Dybala, Rzepka, and Araki (2009a) and Ptaszynski et al. (2009a) in their research on affect analysis of utterances in Japanese. In our task we needed to choose a lexicon proved to be the most appropriate for the Japanese language. Nakamura’s dictionary is a state-of-the-art example of a

<table>
<thead>
<tr>
<th>Emotion categories, dimensions</th>
<th>Language</th>
<th>Approach</th>
<th>Domains</th>
<th>Level</th>
<th>Dataset</th>
<th>Features</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alm et al. (2005)</td>
<td>English</td>
<td>Fusion-based</td>
<td>Fairy tales</td>
<td>Sentence</td>
<td>Children stories (185 tales)</td>
<td>Syntactically, Stylistically</td>
<td>Affect analysis model</td>
</tr>
<tr>
<td>Neviarouskaya et al. (2010)</td>
<td>English</td>
<td>Corpus-based</td>
<td>Fairy tales</td>
<td>Sentence, Document</td>
<td>Children stories (185 tales), News headlines (1890 sentences), Blog entries (700 sentences)</td>
<td>Syntactically</td>
<td>Affect analysis model</td>
</tr>
<tr>
<td>Mohammad (2011)</td>
<td>English</td>
<td>Corpus-based</td>
<td>Fairy tales</td>
<td>Word, Document</td>
<td>192 stories (The Brothers Grimm fairy tales)</td>
<td>Stylistically</td>
<td>Keyword matching</td>
</tr>
</tbody>
</table>
Table 2: Distribution of separate expressions across emotion classes in Nakamura's dictionary, ordered by the number of expressions per class.

<table>
<thead>
<tr>
<th>Emotion class</th>
<th>Number of expressions</th>
<th>Emotion class</th>
<th>Number of expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dislike</td>
<td>532</td>
<td>Fondness</td>
<td>197</td>
</tr>
<tr>
<td>Excitement</td>
<td>269</td>
<td>Fear</td>
<td>147</td>
</tr>
<tr>
<td>Sadness</td>
<td>232</td>
<td>Surprise</td>
<td>129</td>
</tr>
<tr>
<td>Joy</td>
<td>224</td>
<td>Relief</td>
<td>106</td>
</tr>
<tr>
<td>Anger</td>
<td>199</td>
<td>Shame</td>
<td>65</td>
</tr>
<tr>
<td>Sum</td>
<td>2100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hand-crafted lexicon of emotive expressions. It also proposes a classification of emotions that reflects the Japanese language and culture the most appropriately. This classification is also applied in the lexicon itself. All expressions are classified as representing a specific emotion type, one or more if applicable. In particular, Nakamura proposes ten emotion types the most appropriate for the Japanese language and culture. These are: きよろこび (joy, delight; later referred to as joy), こうき (anger), あ ware (sorrow, sadness, gloom; later referred to as sadness), はふく/kowagari (fear), じゃない (shame, shyness, bashfulness; later referred to as shame), かすき (liking, fondness; later referred to as fondness), ないかい/やがる (dislike, detestation; later referred to as dislike), かく/kakaburi (excitement), かややすらぎ (relief), and かよ/やどる (surprise, amazement; later referred to as surprize). The distribution of separate expressions across all emotion classes is represented in Table 2.

A frequent manner in text-based affect analysis research is applying a list of emotion classes based on other modalities than linguistic, such as face recognition, or simply creating a new class list for the need of a particular research see for example comparison of emotion class standards in Ptaszynski, Dybala, Rzepka, & Araki, 2010). The mapping is represented

Table 3: Distribution of separate expressions across emotion classes in Nakamura's dictionary, ordered by the number of expressions per class.

<table>
<thead>
<tr>
<th>Emotion class</th>
<th>Number of expressions</th>
<th>Emotion class</th>
<th>Number of expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dislike</td>
<td>532</td>
<td>Fondness</td>
<td>197</td>
</tr>
<tr>
<td>Excitement</td>
<td>269</td>
<td>Fear</td>
<td>147</td>
</tr>
<tr>
<td>Sadness</td>
<td>232</td>
<td>Surprise</td>
<td>129</td>
</tr>
<tr>
<td>Joy</td>
<td>224</td>
<td>Relief</td>
<td>106</td>
</tr>
<tr>
<td>Anger</td>
<td>199</td>
<td>Shame</td>
<td>65</td>
</tr>
<tr>
<td>Sum</td>
<td>2100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ML-Ask, or Emotive Element and Expression Analysis System is a keyword-based language-dependent system for automatic affect annotation on utterances in Japanese constructed by Ptaszynski et al. (2009). It uses a two-step procedure:

1. Specifying whether an utterance is emotive, and
2. Recognizing the particular emotion types in utterances described as emotive.

ML-Ask is based on the idea of two-part classification of realizations of emotions in language into:

1. Emotive elements or emotemes, which indicate that a sentence is emotive, but do not detail what specific emotions have been expressed. For example, interjections such as "whoa!" or "Oh!" indicate that the speaker (producer of the utterance) has conveyed some emotions. However, it is not possible, basing only on the analysis of those words, to estimate precisely what kind of emotion the speaker conveyed. Ptaszynski et al. include in emotemes such groups as interjections, mimetic expressions, vulgar language and emotive markers. The examples in Japanese are respectively: すげえ (great! – interjection), かわくわく (heart pounding – mimetic), やぐろ (syntactic morpheme used in verb vulgarization) and ！, or ？? (sentence markers indicating emotiveness). Ptaszynski et al. collected and hand-crafted a database of 907 emotemes. A set of features similar to what is defined by Ptaszynski et al. as emotemes has been also applied in other research to discriminate between emotive (emotional/subjective) and non-emotive (neutral/objective) sentences (Wilson & Wiebe, 2005; Wiebe, Wilson, & Cardie, 2005; Aman & Szpakowicz, 2007).

2. Emotive expressions are words or phrases that directly describe emotional states, but could be used to both express one’s emotions and describe the emotion without emotional engagement. This group could be realized by such words as あいおい (love – noun), かんししむ (to feel sad, to grieve – verb), うreshii (happy – adjective), or phrases such as: mushizu ga hashiru (to give one the creeps [of hate]) or ashi ga chi ni tsukanai (walk on air [of happiness]). As the collection of emotive expressions ML-Ask uses the database created on the basis of Nakamura's Emotive Expression Dictionary (Nakamura, 1993).

With these settings ML-Ask was proved to distinguish emotive sentences from non-emotive with a very high accuracy (over 90%) and to annotate affect on utterances with a sufficiently high Precision (85.7% compared to human annotators), and satisfying, although not ideal Recall (54.7%) (Ptaszynski et al., 2009a; Ptaszynski et al., 2009a). To improve the system performance we also implemented Contextual Valence Shifters.

The idea of Contextual Valence Shifters (CVS) was first proposed by Polanyi and Zaenen (2006). They distinguished two kinds of CVS: negations and intensifiers. The group of negations contains words and phrases like “not”, “never”, and “not quite”, which change the valence polarity of the semantic orientation of an evaluative word they are attached to. The group of intensifiers contains words like “very”, “very much”, and “deeply”, which intensify the semantic orientation of an evaluative word. ML-Ask fully incorporates the negation type of CVS with a 108 syntactic negation structures. Examples of CVS negations in Japanese are structures such as: あまり–r (not quite–), ときわる (cannot say it is), or うつ (cannot say it is). As for intensifiers, although ML-Ask does not include them as a separate database, most Japanese intensifiers are included in the emoteme database. The system also calculates emotive value, or emotional intensity of a sentence, on the basis of the number of emotemes in the sentence. The performance of setting the emotive value was evaluated on 84% (Ptaszynski et al., 2009a). Finally, the last distinguishable feature of ML-Ask is implementation of Russell’s two dimensional affect space (Russell, 1980). It assumes that all emotions can be represented in two dimensions: the emotion’s valence or polarity (positive/negative) and activation (activated/deactivated). An example of negative-activated emotion could be “anger”; a positive-deactivated emotion is, e.g., “relief”. The mapping of Nakamura’s emotion types on Russell’s two dimensions proposed by Ptaszynski et al. (2009a) was proved reliable in several research (Ptaszynski et al., 2009a; Ptaszynski, Dybala, Shi, Rzepka, & Araki, 2009b; Ptaszynski, Maciejewski, Dybala, Rzepka, & Araki, 2010). The mapping is represented in Fig. 1. An example of ML-Ask output is represented in Fig. 2.

3.3 Japanese WordNet

WordNet1 (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) is a lexical thesaurus in the form of a database. It was first developed for English in 1990s and later for other languages. Words within WordNet are grouped in synsets, or sets of synonyms representing relations between words. The types of relations include hypernyms, hyponyms, or meronyms. Moreover, adjectives include information on related nouns and other similar words. At present WordNet contains 155,287 words organized in 117,659 synsets. A total number of word-sense pairs is 206,941. As for the Japanese version of WordNet,

---

1 http://wordnet.princeton.edu/
it was developed in 2008 and 2009. It contains 93,834 words organized in 57,238 synsets and 158,058 synset-word pairs. An example of synsets, hypernyms and hyponyms related to word ureshisa ("joy") are represented in Fig. 3.

Apart from the above there exist modifications of standard WordNet, such as WordNet-Affect (Strapparava & Valitutti, 2004) or Senti-WordNet (Baccianella, Esuli, & Sebastiani, 2010), which contain in particular words and synsets applicable in affect and sentiment analysis. Unfortunately, at present they are available only for English, although there has been some research on developing their Japanese versions (Torii, Das, Bandyopadhyay, & Okumura, 2011).

4. Method for affect analysis of characters in narrative

In the task of affect analysis it is crucial to obtain not only the knowledge about the emotions that appeared in the conversation, but also about who expressed them and towards whom/about what they were expressed. This introduces to affect analysis the task of recognizing relations of particular characters with the emotions that appeared.

4.1. Method description

The method consists of three general parts. Firstly, the analyzed text is preprocessed to make the character extraction easier. Secondly, the characters represented by grammatical subject in the sentence are extracted with the use of rules based on combined morphological and syntactic information. Thirdly, affect analysis procedure is performed for sentences in which the subject was identifiable. Below we describe each part of the method. We also compared two methods for affect analysis: the first based on Japanese WordNet (Bond et al., 2009), the second based on ML-Ask system described in Section 3.2. The outline of the method is represented on Fig. 4.

4.1.1. Preprocessing

The preprocessing consists of the following steps.

1. Supplementation of subject in cases of ambiguity (if the subject is omitted).
2. supplementation of subject if the subject is a pronoun.
3. text normalization in cases of ambiguous transcription.
4. preparation of a list of characters appearing in the narrative and character naming unification.

Preparation of the list of characters and subject supplementation are aimed to facilitate the process of character determination. An example of a case covered by 1. including the disambiguation process is represented in the sentence below.

Sentence: なぜかドアが見ると恐怖を感じる(‘‘神’’)
Grammar: Mother TOP home DAT return PAST moment, doors SUB open COP PAST. Afraid become UNIN PAST.
Translation: When mother came back home, the doors were open. She became anxious.

In cases of sentences containing multiple clauses, and in multiple sentences subjects in subsequent clauses are often replaced by pronouns (typical cases of anaphora). In such situations the subject in question is determined as the first previously determined topic (TOP), usually appearing in Japanese before particle “wa”, or subject (SUB), often appearing before particle “ga” (in cases where no topic appeared previously). In this case, the closest previous topic/subject (“mother”) appears in the first sentence. Therefore the default subject for the second sentence is also set as “mother”.

An example of a case covered by 2. including the disambiguation process is represented in the sentence below.

Sentence: お願い wa ie ni kaette kita toki, doa ga aite iruoni ki ga tsuita. Kowaku natte shimatta.
Grammar: Mother TOP home DAT return PAST moment, doors SUB open COP PAST. Afraid become UNIN PAST.
Translation: When mother came back home, the doors were open. [She] became anxious.

2 Grammatical information in accordance to Leipzig glossing rules (Comrie, Haspelmath, & Bickel, 2004). UNIN refers to -shimau, a syntactic marker of non-intentionality and undesirability of action.
Text normalization (point 3.) is performed to facilitate the process of part-of-speech tagging used in character extraction. In Japanese language it mostly refers to normalization of transcription. Japanese is usually written with a mixture of kanji (Chinese ideograms conveying word meaning), hiragana (syllabary used for grammatical information and simple words), and katakana (used for loan words). If most of the sentence is written in hiragana, the syllabograms representing words are normalized to their standard transcription (in either kanji or katakana). An example of a sentence before and after normalization is represented below.

Sentence before normalization: おかあさんはいえにかえってきた。
Sentence after normalization: お母さんは家に帰ってきた。
Sentence in romanized transcription: 仮asan wa ie ni kaettekita toki, doa ga aite ita. Onna wa kowaku natte shimatta.

Grammar: Mother TOP home DAT return PAST.
Translation: Mother came back home.

The preparation of a list of characters appearing in the narrative (point 4.) has a purely practical purpose and for the time cannot be done fully automatically. Many cases, but not all possible, are covered by the below example. Due to the possibility of a mistake we perform a manual correction in this step.

Sentence: 仮asan wa ie ni kaettekita toki, doa ga aite ita. Onna wa kowaku natte shimatta.
Grammar: Mother TOP home DAT return PAST moment, doors SUB open COP PAST. Woman TOP afraid become UNIN PAST.
Translation: When mother came back home, the doors were open. The woman became anxious.

In cases such as above we assume that “woman” refers to “mother”.

4.1.2. Story character extraction

From the narrative preprocessed as in Section 4.1.1, sentences containing indications of characters of the narrative are selected and the characters are specified. To do this firstly the sentences are tagged with parts of speech (POS) by a Japanese POS tagger MeCab (Kudo, 2001) and parsed by a Japanese dependency parser CaboCha (Kudo & Matsumoto, 2002). The information provided by these two systems is applied similarly to the algorithm proposed by Ogi and Suganuma (2009) for the detection of topics/subjects and predicates in sentences:

(i) Consider the clause ending with a period as the clause containing the main predicate.
(ii) Consider clauses related to the predicate clause as potentially containing candidates for subject clause.
(iii) Within the clause with subject candidates find particles wa (topic marker) and ga (subject particle).

(iv) From the candidates appearing before wa and ga consider those related to the predicate clause as the subject/topic.

In the preprocessing phase all sentences are normalized and subjects are supplemented. Moreover, the list of narrative characters is prepared. A cross-reference of clauses containing topic/subject with the list of characters allows extraction of only those sentences containing a character of the narrative as the topic/subject of the sentence.

4.1.3. Affect analysis

In this section we describe two methods for affect analysis. The first one based on Japanese WordNet (baseline) and the second one based on ML-Ask.

WordNet (Baseline). The procedure for applying Japanese WordNet in affect analysis is as follows. The sentence is tagged with parts of speech and from each sentence nouns, verbs and adjectives are extracted for further analysis. All those words are separately queried in Japanese WordNet. If the word relates to emotions on any level or hypernym of hyponym, it is extracted as a candidate for emotion classification. As an example, hypernyms of word kanashimi (“sadness, unhappiness”) are represented in Fig. 5. The word kanashimi is included in many synsets and thus multiple results can be obtained for this word. For example, gradual hypernyms of kanashimi are kan (“feeling”), dosei (“state”), atoribyō (“attribute”) and shashō (abstract_entity). For “sorrowfulness” there is an hypernym fuyo (“unhappiness”), and for “gloominess” there are such hypernyms as kikigen (“uncheerfulness”), atoribyō (“attribute”) and shashō (abstract_entity).

In particular, in case of nouns we assume that if within synsets appear hypernyms/hyponyms like jōkan (“emotion”) and kan (“feeling”) the word is emotion-related. We assume that a verb is emotion-related if among its synsets appear emotive verbs like kanjiru (“to feel”) or kanjiru (“to experience”). In case of adjectives, the computation is somewhat different. Adjectives themselves do not contain synsets like hypernyms or hyponyms. However, they contain nouns and verbs related to those adjectives. Therefore we perform an extrapolation to the synsets of those related words to check if these relate to emotions (appearance of jōkan (emotion), kan (feeling), or kanjiru (to feel/experience)). If they do we assume the adjectives are emotion-related as well.

ML-Ask simple. As the second method we used the information obtained by analysis of sentences by ML-Ask. However, some modifications needed to be made to the main system. ML-Ask was designed to analyze conversations, not narratives. In the first step of ML-Ask analysis the system specifies if a sentence is emotive or non-emotive. Analysis of particular emotion types is performed only on emotive sentences. A sentence is emotive if it contains at least one emoteme, or a marker of emotive context. Emotemes are typical in conversations (in particular spontaneous conversations). Generally perceived narratives contain at least two main types of sentences:
1. **descriptive sentences** for introduction of the main storyline, and
2. **dialogs** between characters of the narrative.

ML-Ask baseline can be expected to deal with the second type of sentence. However, since emotemes rarely appear in descriptive sentences, the system would not proceed to the recognition of particular emotion types. Therefore, to allow ML-Ask deal with descriptive sentences as well we excluded emotemes from the analysis and focused primarily on analysis of emotion types. However, we retained the analysis of CVS and Russell’s emotion space. Since in this version of the system we simplified the analysis, we called it **ML-Ask-simple** (later abbreviated to ML-Ask).

5. **Evaluation experiment**

5.1. **Dataset**

To evaluate the method we needed to choose a narrative. A popular kind of narratives applied in affect analysis are children stories and fairy tales (Alm et al., 2005; Mohammad, 2011). This kind of narrative usually has a clear structure and distinction between good and evil characters. Fairy tales are written in a language simple enough to be understood by young children, with comprehensible descriptions of emotions felt by the characters. Therefore we decided to evaluate the method on children stories. As the source of the dataset we used **Aozora Bunko**.

_Aozora Bunko_ ("Blue Sky Library") is a Japanese online digital collection of freely available books (online library). It provides over ten thousand books of various genres (fiction, non-fiction) published in Japanese, for which copyrights had expired (50 years since the death of copyrights holder). In Japan, _Aozora Bunko_ is sometimes compared to Project Gutenberg (Tamura, 2006).

One of the genres in _Aozora Bunko_ is _Jidōsha_ ("Books for children"). This category includes sub-categories, such as _Rekishi_ ("History Books"), _Geijutsu-Bijutsu_ ("Beautiful Arts and Crafts Books"), and _Bungaku_ ("Literature"). Children stories and fairy tales are included in the _Bungaku_ category. It covers 843 books for the date (March 2012). From this database we chose one fairy tale at random for the experiment.

The outline of the story can be presented in the following eleven parts.

1. The Goat Mother leaves the Goat Children home alone to look after the house. Before leaving the Mother alerts the children to be on their guard for the wolf. The children reply they will take care of themselves.
2. The Wolf comes to the Goats’ house. He pretends to be the Mother by changing his voice, but the Children recognize the Wolf by the rough voice.
3. The Wolf changes his voice to sound more softly (by eating a chalk) and comes again to the Goats’ house. Again the Children recognize him by the color of his foot claws.
4. The Wolf bleaches his foot by putting some flour on it. The Children believe the Wolf and open the doors.
5. The Wolf eats all Children Goats except one. After that he falls asleep under a tree.
6. Mother comes back home. She looks for her Children in the devastated house.
7. The Mother listens to the explanation by the one Child Goat that was not eaten by the Wolf.
8. The Mother finds the Wolf, cuts his stomach with scissors and takes out her Children.
9. Mother quiets down the Children jumping of joy, and tells them to bring some stones.
10. While the Wolf is sleeping the Mother puts the stones in his stomach and sews it together. Then she flees from the scene. The Wolf wakes up, but he cannot move because of the stones.

11. The Wolf attempts to drink water from a well, but falls inside. The Goat Mother together with the Goat Children cries out “The wolf is dead! The wolf is dead!” together they dance for joy around the well.

A story divided roughly in these 11 parts was presented to nine evaluators (all of them were graduate students). The evaluators were to read the story and annotate emotions for each of the character in each of the eleven parts of the story. The evaluators were encouraged to annotate emotions spontaneously with simple and instinctive vocabulary, like “joy”, “sadness”, or “anger”. The annotations were then automatically classified according to the emotion classification applied in the research. In case of no emotions the evaluators were told to annotate “none” on the person/story part. This constituted the gold standard.

The story was then processed as a whole by the system including two versions of affect analysis (WordNet and ML-Ask). After the processing the extracted and analyzed characters were allocated accordingly to their appearance in all 11 parts of the story. The results of the system were then compared to the gold standard. All annotations are represented in Table 3.

5.2. Evaluation criteria

We evaluated whether the system output was equal to the annotations set by the evaluators. However, since there were some cases of disagreement about the emotions, in the evaluation we used only those types for which more than 50% of the people agreed (at least 5 people). The percentage of agreement, or agreement ratio is represented in Table 3 next to the emotion type annotated by evaluators. For some cases the agreement was 100%. In cases where not all evaluators agreed, we did not include the outliers’ annotations. However, even in this setting there were three ambiguous cases. For the Goat Children in part 4 of the story, there were two most agreed emotions: anger (33%) and joy (33%). For the Wolf there were two cases of ambiguity, in part 2 of the story three people (33%) agreed about anger and three about the lack of emotions. Similarly, in part 7 four people (44%) agreed on joy and other four on the lack of emotions. In the evaluation we verified whether the ambiguity improved or impaired the analysis. We also looked at the results in two different ways. Firstly, since the goal was to annotate emotions, we assumed that cases of no emotions are irrelevant and excluded them from the evaluation. Secondly, since both, the human evaluators and the systems had the option of “no emotion”, we checked the performance of the systems when the cases of no emotion were included in the evaluation. The accuracy of the systems’ performance was calculated as a standard ratio of successful cases to all gold standard cases, as represented in Eq. (1).

\[
\text{Accuracy} = \frac{\text{successful annotations}}{\text{gold standard annotations}} \tag{1}
\]

Accuracies of all versions of the system are represented in Table 4. For ML-Ask we also checked two options. Firstly, when the annotated emotion types are exactly the same as those in gold standard (types-strict). Secondly, when the annotated emotion types, if not exactly the same as in gold standard, at least share the same Russell’s space (types-relaxed). We also calculated statistical significance for the differences between WordNet-based method and ML-Ask.

5.3. Results

The results were as follows. For most cases the method based on ML-Ask was better than the WordNet-based method. The results were somewhat close, with 0.576 as the best result for WordNet and 0.606 as the best result for ML-Ask (difference equal to 0.03). Statistical significance of differences was confirmed for two out of three sets of results. The results were very statistically significant (p < 0.01) for the Goat Mother and extremely statistically significant (p < 0.001) for the Goat Children, but not quite statistically significant for the Wolf. However, detailed analysis of the results revealed that WordNet method was able to extract only emotions from the class of “joy”, while ML-Ask was able to extract various emotion types, including “joy”, “fondness”, “relief”, “feel”, “sadness”, or “anger”. Therefore it can be said that ML-Ask was more balanced in affect analysis, while WordNet was more biased (towards expressions of “joy”) when it comes to determination of emotion types.

Analysis with the type “no emotion” included showed better performance in both methods, which means that the system, although not always extracting emotions when it should, rarely extracts emotions from sentences that are not emotional. As for the two options of ML-Ask evaluation, where the extracted types are considered as strict or relaxed, the relaxed method achieved the best result of all. Moreover, the results compared to the strict evaluation were statistically significant on a standard 5% level. Detailed analysis of errors showed interesting dependencies. In scene 9, where the Mother quiets down the Children jumping of joy, and tells them to bring some stones, the Mother Goat was annotated by most of human evaluators (70%) as filled with anger. Both systems however annotated “joy”. This was due to the fact that although “joy” is expressed here directly, and the mother does feel happy because the Children are safe, the dominating emotion here is anger (for Children to be quiet and for the Wolf in general). However, the information about the anger is not present directly in any form. It is inferred indirectly by the situation (context). For hu-
Table 4
The summarized results for all kinds of evaluation criteria.

<table>
<thead>
<tr>
<th></th>
<th>Goat Mother</th>
<th>Goat Children</th>
<th>Wolf</th>
<th>Overall/average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WordNet</td>
<td>ML-Ask</td>
<td>WordNet</td>
<td>ML-Ask</td>
</tr>
<tr>
<td></td>
<td>Types-strict</td>
<td>Types-relaxed</td>
<td>Types-strict</td>
<td>Types-relaxed</td>
</tr>
<tr>
<td>With non-emo</td>
<td>0.636</td>
<td>0.727</td>
<td>0.564</td>
<td>0.565</td>
</tr>
<tr>
<td>Only emo</td>
<td>0.633</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>Partial average</td>
<td>0.485</td>
<td>0.614</td>
<td>0.576</td>
<td>0.576</td>
</tr>
<tr>
<td>Without ambiguous sentences</td>
<td>0.636</td>
<td>0.727</td>
<td>0.564</td>
<td>0.565</td>
</tr>
<tr>
<td>Only emo</td>
<td>0.633</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>Partial average</td>
<td>0.485</td>
<td>0.614</td>
<td>0.576</td>
<td>0.576</td>
</tr>
</tbody>
</table>

*p < 0.05.
*<sup>a</sup> p < 0.01.
*<sup>b</sup> p < 0.001; with comparison to baseline (WordNet).

manns this ability is inherent. The mechanism of this reasoning should be studied deeply in the future.

The evaluation of particular emotion types provided by evaluators (gold standard) along with all annotations given by both systems are represented in Table 3. The results for all kinds of evaluation criteria are summarized in Table 4.

6. Conclusions and future work

In this paper we presented our research in text-based affect analysis (AA) of narratives. Affect analysis within the context of narratives is a challenging task for several reasons. Firstly, different types of sentences are used in narratives (descriptions, dialogues, etc.). Secondly, different characters become subjects of different emotional expressions in different parts of narratives. In this research we addressed in particular the problem of person/character related affect recognition in narratives. In this method firstly emotion subject is extracted from a sentence based on analysis of anaphoric expressions. Next, the affect analysis procedure estimates what kind of emotional state each character is in for each part of the narrative. We compared two methods for affect analysis. The first method was based on Japanese version of WordNet, a thesaurus consisting of words with specified relations to each other. The second method was based on ML-Ask, a system for affect analysis of Japanese language. The evaluation showed that the ML-Ask achieved significantly higher results. Although the results in general are not ideal (60.6%), they were still promising as this was the first attempt to combine two tasks: affect analysis in narratives and story characters.

We plan to improve the method in several ways. Firstly, the affect analysis system can be improved by expanding its emotion lexicon. We plan to utilize Ptaszynski et al.'s SPEC system (Ptaszynski, Rzepka, Araki, & Momouchi, 2010) to expand the number of syntactic patterns in ML-Ask. Secondly, the affect analysis procedure can be improved by applying a Web-mining technique for emotion association extraction, which showed an improvement in the past (Ptaszynski et al., 2009a).

Affect analysis of narratives is an important task crucial in language understanding. The ability to distinguish which character is in what emotional state at a given moment is especially important in this task. We showed this task is feasible to some extent for the Japanese language with available tools. Other important ones include the ability to determine why one expresses the emotion (determination of emotion object), ability to distinguish whether the expression of emotion is appropriate to the situation, and what would be the appropriate reaction. In this research we focused on the first task on the list, however, we plan to gradually address all of the above tasks to contribute to deeper understanding of emotion-related phenomena.

Acknowledgements

This research was supported in part by JSPS KAKENHI Grant-in-Aid for JSPS Fellows (No. 22-00358) and a Grant-in-Aid for Scientific Research (No. 24500601).

References


