Automatic character coding in dream reports

Douwe van Erp
Radboud University Nijmegen, Netherlands
douwe.erp@student.ru.nl

ABSTRACT

This paper explores the possibility of automating the character coding process of the Hall-Van de Castle system for dream annotation. We propose a rule-based system that utilizes WordNet, named entity recognition and coreference resolution. The reliability of this system is evaluated by the percentage-of-agreement method. Although the system's reliability is not yet good enough for practical purposes, the results are still very promising and suggest that further development could lead to a reliable automatic coding system.

KEYWORDS

natural language processing, entity extraction, wordnet, dreams, content analysis

1 INTRODUCTION

The field that is occupied with the rigorous study of dream content is known as quantitative dream analysis. Central to their methodology is the Hall-van de Castle coding system [6]. The system was developed and refined in order to categorize a textual dream report into several nominal and clearly defined categories, resulting in a high intercoder reliability [2]. Consequently this system allowed researchers to obtain statistical norms for dream content, and further allows for statistical analysis of dreams in general.

Coding dream reports into the Hall-Van de Castle coding specification has traditionally always required two independent human annotators. Annotations are made based on several categories, such as the characters in a dream, their social interactions, their activities, their misfortunes and good fortunes, their emotions and the dream setting.

However, given the past few decades of advances in the machine learning and natural language processing it is not inconceivable that this coding process could now be reliably automated. Because this has not been attempted before, it could prove to be a great advancement for both quantitative dream analysis and content analysis. This paper sets out to examine the possibility of automatically annotating characters in dream reports. Character coding is the required first step to automate the full coding process, since the other categories make use of character codings. Accordingly, we propose a rule-based system that treats character coding primarily as an entity extraction problem.

In section 2 we take a look at related work regarding entity extraction. In section 3 we describe the data-set for testing, the character coding specification and our system design. In section 4 we assess the reliability of our system, analyze the contributions of our individual system components, and perform failure analysis to get insight in our system's shortcomings. Finally in section 5 we discuss the results, along with possible follow-ups for this line of research.

Copyright held by the owner/author(s).
TxMM'17-'18, Radboud University, Nijmegen, Netherlands

2 RELATED WORK

At first glance this seems like quite unique work, because it needs to combine solutions to different kinds of natural language processing problems.

There has been some investigation in the area of automatically coding dreams, but this is mostly limited to annotating emotion or sentiment or in dream reports. [5] explores automatic emotion annotations of dream reports, for which they use the Hall-Van de Castle emotion coding categories. However their goal was not to automatically duplicate coded dreams, since they also include implicit emotional markers, and do not associate emotions with dream characters.

Our entity extraction problem is mostly related to animacy classification, as we need to extract all persons and animals from the text (see 3.2). Animacy classification allows for instance the annotation of noun phrases according to fine-grained semantic hierarchies, i.e. with a maximum entropy classifier [1]. Another very powerful approach is to exploit the animacy information from the word senses in *WordNet*, which can be done in a rule-based manner as well as with machine learning [4].

3 APPROACH

The following subsections describe our data-set, the coding specification that our system must conform to, and consequently an implementation for a rule-based system that can perform automatic coding.

3.1 Data-set

To assess the reliability of our system, we test it on the Hall-Van de Castle norm dreams. These consist of 500 male dreams and 500 female dreams, recorded from 100 males and 100 females (5 dreams per person) at Case Western Reserve University. The samples and their corresponding Hall-Van de Castle codings have been obtained from *DreamBank*[1]. By excluding the dreams with missing reports or without character codings, this results in 936 samples.

One shortcoming is that the dataset only lists the character codes for a given dream report, but not the specific part of the text from which a character is extracted. This precludes direct use of supervised learning approaches, i.e. learning the correspondence between a partial phrase and a character code.

An advantage is that the reports are clearly formatted and contain little syntactical and grammatical errors, which is beneficial to accurate sentence parsing.

[1] http://www.dreambank.net/search.html

3.2 Character coding specification

The characters in dream reports are annotated according to the rules of the Hall-Van de Castle system [2, 6]. A thorough explanation has been outlined by DreamResearch [8], of which we will give a condensed overview. However, some additional rules are left out in order to limit the scope of this project.

Characters consist of people, animals and mythical figures that are present in the dream. The dreamer is not coded because this would be redundant. Furthermore, characters are classified according to four classes, each designated by a letter or number. Thus a character is always coded with a four-symbol code, in the order Number-Gender-Identity-Age.

3.2.1 Number. Number makes a distinction whether the character is a single individual or a group. An individual (e.g. *the man*, *my professor*, *a dog*) is coded by `1` and a group (e.g. *the crowd*, *several people*, *seven dwarfs*) by `2`.

Animals are also classified as individuals or groups, but not in terms of gender, identity and age. Animals are assigned `1ANI` for individuals, and `2ANI` for a group.

Numbers `3` to `8` are used to describe characters that are dead, imaginary or have changed form. Since they have a relatively low occurrence, they are not considered for our initial system.

3.2.2 Gender. The first two sub-classes denote male (`M`) and female (`F`). Additionally, groups that are made up of both males and females (e.g. *parents*, *the classroom*) have a joint gender (`J`). If the gender can not be inferred from the context it is classified as indefinite (`I`).

3.2.3 Identity. Identity consists of sub-classes that indicate a character's familiarity to the dreamer. Immediate family includes: *father* (`F`), *mother* (`M`), *parents* (`X`), *brother* (`B`), *sister* (`T`), *husband* (`H`), *wife* (`W`), *son* (`A`), *daughter* (`D`), *child* (`C`), *infant* (`I`) and *baby* (`B`). Relatives (e.g. *aunt*, *cousin*, *grandpa*) are denoted by `R`. Other characters are known (`K`) if the dreamer is personally acquainted with them. Next are prominent persons, e.g. *Shakespeare*, (`P`), characters that are only occupationally (`O`) or ethnically (`E`) identified, and strangers (`S`). If the familiarity of a character to the dreamer is unknown, it is designated with `U`.

3.2.4 Age. A character is seen as adult (`A`) if it does not fit in another age group. Characters are classified as teenagers `T` if their age is between 13-17, or the context shows that they are an adolescent. Characters from age 1-12 are children (`C`), and characters below age 1 are babies (`B`).

3.3 System design

What follows is a rule-based system based on the requirements from 3.2. It consists of several components that solve tasks related to five different categories.

3.3.1 Entity extraction. The first step is to extract entities from the text that are potential dream characters. My first approach was with statistical named entity recognition (NER), but this only finds (well-known) person names in the text. Unnamed entities like *the man* and *a large crowd* also need to be recognized. To solve this we use the observation that entities almost always are referred to by noun phrases, which we can treat as *candidate* characters. For each report we apply tokenization, POS-tagging and dependency parsing with *spaCy*[2], in order to iterate over noun chunks.

The next challenge is to detect whether a noun phrase refers to an animate entity or an inanimate entity. Stanford's *CoreNLP*[3] offers animacy annotation in their co-referencing component, but in practice this yielded inaccurate results. Instead our solution utilizes *WordNet* [7], a lexical database that groups English nouns into sets of synonyms (synsets) and interlinks them with other synsets based on their semantic and lexical relations.

First we query the synsets for the root noun of a noun chunk. This return all synonyms for that noun, and thus also helps us to deal with ambiguous nouns (e.g. *class* can refer to a category, but also to a classroom of people). For every synset we traverse their hypernym taxonomy to find the semantic categories they belong to. If a synonym of the noun belongs to the hypernym `person` or `animal`, we select it as a character. Similarly we can extract mythical creatures with the hypernym `imaginary_being`, but for now we disregard them because of their low occurrences in dream reports.

Initially we will assume the code `1IUA` (individual, indefinite gender, unknown identity, adult) and adjust it as more information about the character is acquired.

3.3.2 Number identification. To determine the number of a character we will first look at whether it has a singular or plural word form. For instance when we query *WordNet* for the noun *women*, this will resolve to the lemmatized form *woman*. Therefore if the noun is not equal to the returned synset, but the lemma of the noun is in fact equal to the returned synset, this indicates plurality.

Other cases like *group* can not be classified with this rule, because both the singular and plural form refer to a group (or groups) of individuals. These cases are detected with a second rule that checks if the noun belongs to the hypernym `people`.

3.3.3 Gender identification. To identify a large group of nouns we can apply a simple gender test by looking at the suffix of the noun. Nouns ending in *-man*, *-men* and *boy(s)*, are coded as male (M). Similarly, nouns ending in *-woman*, *-women*, *girl(s)* and *-maid(en)(s)* are coded as female (F).

Sometimes the gender can not extracted from the noun itself, but it can be inferred from third-person pronouns that follow it. For instance in report #3 of the male norms: *"I was talking to my friend MB, 22, at my house.* ***He*** *said something..."* we can determine that the gender of *MB* is male. Based on this we implement a very basic co-referencing technique that checks if a noun chunk is a third-person pronoun. If this is the case, we adjust the gender of the *last* found character accordingly.

3.3.4 Identity identification. Identifying the family members from and relatives is partly trivial. For this we use dictionary NER, and essentially look-up if a noun matches the name of a family member or relative. In most cases this also gives us information about the number, gender and age.

Additional identifications are all based on *WordNet* hypernyms that were found to be inclusive. We determine occupation (`O`) by the hypernyms `professional`, `employee` , `expert`, `organization`

[2] https://spacy.io/
[3] https://stanfordnlp.github.io/CoreNLP/

Table 1: Intercoder reliability by percentage-of-agreement.

Category	Percentage-of-agreement between automatic coder and human coder	Percentage-of-agreement between human coders [3]
Characters		
Presence	59	93
Number	92	92
Gender	68	89
Identity	48	81
Age	93	-
All correct	28	76

and `skilled_worker`. Otherwise known characters (K) typically can be determined by `neighbor`, `peer`, `acquaintance`, `relative`, `kin`, `leader` and `friend`. For now we ignore the ethnic subclass (E) because it is uncommon, and also the stranger subclass (S) as it more difficult to determine and should be inferred from the context.

Finally we also include *spaCy*'s statistical NER to find prominent persons (P) that are not always extracted by our hypernym rules. For instance in dream report #6 of the male norms: *"...Then the voice said, "that is Beethoven's 4th. How could you ever forget."..."* this will extract *Beethoven* as a prominent person, whereas it was otherwise not found.

3.3.5 Age identification. The ages of our dreamers are all adult (A), therefore we assume most of the familiar characters of the dreamer (friends, cousins, siblings) also to be adult. Otherwise occupational identities and prominent persons are also assumed to be adult.

As an extra age test, we say that the teenager (T), child (C) and baby (B) subclasses are respectively determined by the hypernyms `adolescent`, `child` and `baby`.

4 RESULTS AND ANALYSIS

4.1 Reliability

We determine the reliability of our system by the percentage-of-agreement method, which is typically how two human Hall-Van de Castle coders are compared [2]. This can be calculated as

$$p = \frac{\#agreements}{\#agreements + \#disagreements} \approx \kappa$$

Because there are 2432 possible character codings, the hypothetical change agreement is very small. Therefore the percentage-of-agreement is also approximately equal to Cohen's κ.

Table 1 shows the agreement percentages for each class of the character category. Additionally it also shows the agreement about the presence of the number of characters, and the percentage of characters that are perfectly agreed on.

Table 2 shows the number of dreams coded, and the number of character codings by our system (coder A) and by the dataset (coder B). When a different number of characters was found between coders, we tried the possible permutations to find the best match. 25 dreams were removed due to a high number of characters, which resulted in a high number of possible permutations.

Table 2: Number of dreams and characters coded. For Automatic coding, coder A is automatic and coder B is human. For human coding, both coder A and B are human.

Number of	Automatic coding	Human coding [3]
Dream reports	911	100
Codings by A	3892	276
Codings by B	2450	276

Table 3: Decrease in percentage-of-agreement when leaving out a component of the system.

Component	Number	Gender	Identity	Age	All
Extraction (3.3.1)					
Animal	0	2	4	2	2
Number (3.3.2)					
Plurality test	**12**	1	-1	0	2
Group	2	1	0	1	1
Gender (3.3.3)					
Gender test	0	**11**	0	0	2
Coreference	0	**8**	0	0	3
Identity (3.3.4)					
Dictionary NER	0	**8**	**7**	0	**8**
Occupation	0	-1	4	0	1
Known test	-1	-2	**14**	0	**5**
Prominent NER	0	1	0	0	0
Age (3.3.5)					
Age test	0	0	0	1	0

Both tables also show a side-by-side comparison to the results of the reliability study between human coders from [3].

4.2 Component contributions

We assessed how much certain system components contributed to the reliability of the system. Table 3 shows the decrease in percentage-of-agreement when a component was left out. Negative percentages denote an increase in the percentage-of-agreement.

4.3 Failure analysis

To gain insights in what type of errors the system makes, we have manually inspected the results of the 10 first male norm dream reports (see appendix). Based on this we have identified several categories of errors.

The first category are not really system mistakes, but are based on disagreement with the human coder. In dream #1, #2, #5 and #9, the respective characters *the professor*, *a friend of mine*, *several fellows* and *some hunters* are all coded as male (M) while this cannot be inferred from the context. The system however correctly codes them as indefinite I and joint J.

The second category is based on missing coreferences. For instance in #3 *his uncle* is coded as a relative (R), while the possessive pronoun indicates that it refers to the uncle of a familiar person and is therefore known (K). In #5 *a model* is coded, but refers to the already coded character *a girl my age*. In #7 *the persons* and *the*

man are coded, but refer to the previously coded *a woman* and *her neighbor*.

The third category mistake is identifying superfluous characters. This results from checking the hypernym taxonomies for all possible synonyms of an extracted noun. In #6, *the rail* can refer to the railway organization (group), *a voice* is a synonym for a singer and *a name* is a synonym for an important person. In #8, *blues* is a synonym for a type of buttery, and is thus seen as a group of animals. In #9 and #10, *snow*, *tree* and *woods* are picked up because they exist in *WordNet* as last names for writers and actors.

Moreover there are also more incidental mistakes. In #2, *some people that I know* are coded as unknown (U) instead of known (K), because the relative clause modifier *that I know* is not taken into account. In #1, *the professor* is coded as occupational. This is not false, but the context indicates that the character is related to the dreamer, and should thus rather be classified as known (K). Similarly in #8 the context also needs to be understood in order to code *persons* as stranger (S). A very challenging error occurs in #10: *"...I woke up my wife"* refers to a commentary of the dreamer outside of the dream. However the system does not see this distinction and incorrectly extracts *my wife* as a character.

The main source of error seems to be in entity extraction, followed by wrong character identification based on missing coreference information. Moreover, incidental mistakes primarily fail to incorporate context information.

5 DISCUSSION AND OUTLOOK

Considering that the system was started from scratch I think that it performs surprisingly well. In total it correctly extracted and coded 28% of the characters that were also recognized by the human coder. It is however not yet useful in practice, for which the percentage-of-agreement should be around 76% (which allows for a margin for disagreement). Interestingly, the agreement on the number class of the character already corresponds with the reliability of human coders. The agreement on age is also very high, but this could be due to a statistical bias because we initially assume a character to be adult, which holds up for most dream characters.

The main shortcoming of the current system seems to be that too many characters are extracted (3892 by the system as opposed to 2450 by the human coder). This may give the system an unfair advantage as it can try to match more of its coded characters with the characters of the human coder. Ideally we should constrain the number of character such that it is closer to the number identified by the human coder. One way to do this could be to select only one synonym for a hypernym search, instead of trying out all synonyms. For this we need to which synonym best fits in the context of the dream. A possibility is to sum the (semantic) similarity between the synonym and each other word in the sentence, for instance by *WordNet*'s *lch* measure. The synonym with the highest similarity to the sentence is then likely to be the right fit for the context.

The component analysis shows us that almost all components increase the reliability of the class for which they are designed. Especially the 14% increase in the identity agreement by the *known test* shows how powerful utilizing *WordNet* hypernyms can be. The only exception is the prominent person NER, which does not give an increase for the identity class. Since our very simple coreferecing technique already gives an 8% increase in gender identification agreement, more advanced coreferencing could potentially improve the system's reliability. More advanced coreferencing could also solve some of the third category errors that we identified in failure analysis (4.3).

To conclude: despite the large scope of this project, the preliminary results already show promise that the goal of automating character coding could in fact be achieved. However the current system is yet far from practical and should thus be refined. Possible follow-ups with supervised learning methods could also prove useful. Although these might not have the same consistency of rule based systems, and often behave as a *black box* with regards to their inner workings. Ultimately, the follow-up to successful automatic character coding would be to automate the rest of the Hall-Van de Castle coding process (i.e. social interactions, activities, emotions, etc). Character coding is however at the very basis of this, which therefore requires it to be highly reliable.

ACKNOWLEDGMENTS

Gratitude to William Domhoff and Adam Schneider for providing me with the DreamBank Coding Search Utility and for words of encouragement.

REFERENCES

[1] Samuel R Bowman and Harshit Chopra. 2012. Automatic animacy classification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Association for Computational Linguistics, 7–10.

[2] G William Domhoff. 2003. *The scientific study of dreams: Neural networks, cognitive development, and content analysis.* American Psychological Association.

[3] G William Domhoff. 2013. *Finding meaning in dreams: A quantitative approach.* Springer Science & Business Media.

[4] RJ Evans and Constantin Orasan. 2007. NP Animacy Identification for Anaphora Resolution. *Journal Of Artificial Intelligence Research* (2007).

[5] Elena Frantova and Sabine Bergler. 2009. Automatic emotion annotation of dream diaries. In *Proceedings of the analyzing social media to represent collective knowledge workshop at K-CAP 2009, The fifth international conference on knowledge capture.*

[6] Calvin S Hall and Robert L Van de Castle. 1966. The content analysis of dreams. (1966).

[7] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[8] Domhoff G. W. Schneider, A. 2018. The Classification and Coding of Characters. (2018). http://www2.ucsc.edu/dreams/Coding/characters.html Accessed: 2018-01-19.

APPENDIX

Included are the first 10 dreams of the Hall-Van de Castle male norm dreams to accompany the failure analysis in 4.3.

#0001 *I was in professor Teimes' classroom in Corp. Finance. The room was larger than usual. We had several women in the class, which is not as it is. The professor was asking the class questions on our work. Being unprepared, I feared the time when I would be asked a question. Finally I was asked a question, which evidently I answered*

all right, since there were none of the repercussions I had feared due to my unpreparedness.

#0002 *I was a large reception room which was given by some people that I know. I seemed to be busily at work gorging myself with all types of foods. I noticed a huge display of drinks which was actually amazing in that the containers which were orange and red and yellow were piled on top of each other in a pyramid fashion. I believe a friend of mine entered the picture, and then I woke up.*

#0003 *I was talking to my friend MB, 22, at my house. He said something about his uncle getting a new British car for $8000. The next thing I know, I was outside heading for the car when I saw a little puppy who had been and was throwing up. I seemed to shake my head and say, too bad. I guess he's got distemper. I then proceeded to inspect this super streamlined car which was a convertible and was parked across the street.*

#0004 *There were three men in our recreation room whom I recognized but could not remember when I woke up. I went down the stairs and was greeted by a burst of pistol fire. Three slugs ripped the lower right part of my stomach. I was wiping the blood on my shirt as I walked back up the stairs. I then seem to remember being with a dog who was in a similar condition. It seemed very realistic, and I was in great pain until I woke up feeling my right side and was relieved when I found myself in good shape.*

#0005 *I was in a gym with several fellows, including an instructor at reserve who said his name was La Rien. Actually his name is quite different. We were all running around the track above the gym. Suddenly a girl my age who is now married and is also a model is seen swinging on one of the ropes which are suspended from the ceiling. She swings back and forth and occasionally grabs onto the rail next to the track.*

#0006 *I was walking all alone down what must have been a quiet street, as I remember no traffic noise of any sort. As I passed one rather large house, music floated out to me, and I became quite angry because I could not think of the name of the piece. It began to rain then, but instead of hurrying on my way I stood and listened to the music. I was just about ready to give up when a low soft voice said "don't you remember that, my dear." I turned, but no one was in sight. Then the voice said, "that is Beethoven's 4th. How could you ever forget." I looked frantically around but could see no more. Something made me look at the house for a moment, and for a second I saw a lovely face watching me. She seemed to be crying, yet I wasn't sure. The rain came down harder, the music ceased, and I walked on alone.*

#0007 *This dream was unusual for me as the time and place seem to be absent. I seem to have been walking along a quiet street, and suddenly coming upon a woman screaming at her neighbor because smoke from his fire was getting her wash dirty. The persons were not clear, so I can not identify them. The black smoke and white wash seem to stand out. The thing I remember most is how funny it seemed to me at the time to see the woman yelling at the man and pointing to her ruined wash.*

#0008 *This dream seems to have taken place in a hotel ball room. It was graduation night, and my wife and I were celebrating by having our big night out. The room and people were very real and colorful. The one impression I had upon awakening was that everything seemed so real and alive. The women's formals I remember as bright red's, greens, and blues. The good was just as tasty as it is real, and I remember how satisfied I felt after we ate.*

#0009 *I was hunting, all alone in what seemed to be a thick wooded area. I was carefully following deer tracks in the snow. They were fresh, as ice had not yet formed in the print, and I was very excited. All of a sudden, there it was, at least an eight pointer. I took aim and fired. He fell, but scrambled up at once and hurried off. I gave him a good half hour start and then followed. In an hour or so, I found him bleeding in a little gully. But now as I looked at him trying so desperately to get up and run away, I became sick to my stomach and ashamed of what I had done. Just then some other hunters came up, praised me for my prize, and made quite a fuss over the buck. I gave the kill to them and left.*

#0010 *This dream seemed to be in a forest, although I do not know exactly where. I was sitting on a log apparently waiting for a squirrel to come into sight. The birds were singing, and I felt very sleepy and warm. The squirrel came out on a branch high up in a tree, and I took careful aim but could not fire. I must have said "run and play little fella" rather loud because it woke my wife up who in turn woke me. I was conscious of the green moss and little red flowers in the woods as well as the beautiful reddish coat of the squirrel.*