Modelling human appreciation of machine generated *What-if* ideas

Jasmina Smailović, Martin Žnidaršič, Nada Lavrač, Senja Pollak, Janez Kranjc
Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
{jasmina.smailovic,martin.znidarsic}@ijs.si

Abstract

In this paper we present our attempt at modelling the human appreciation of fictional ideas. The appreciation models are an essential part of the automated generative loop of these artefacts, which can be used in story-writing, advertisement and similar domains. The ideas in our case are machine generated by computational creativity systems and are represented in textual form. We present our approach to machine learning of such models and the results of their empirical assessment.

1 Introduction

Fictional ideas are propositions of situations that are unrealistic or commonly considered as unplausible, such as:

What if aliens from another planet came to Earth for holidays?

which are commonly a central part of various creative works and products. Understanding and modelling of human evaluation of fictional ideas is one of the aims of the EU project *The What-If Machine (WHIM)*[1], which studies computational production of such ideas. In the context of WHIM and this work, the fictional ideas are represented in textual form as sentences that start with "What if " and propose a fictional situation (see Table 1 for some examples). We will refer to them as the *What-if* sentences or *What-ifs*. Artificial production of *What-if* ideas is creative work that is inherently hard to automate, but there are now some generators available [3, 4, 8]. However, their results often suffer either from a narrow covering of the *What-if* idea space, either from a very low quality of the final results. Sensitivity to one or the other problem is usually a trade-off between a more template driven or more open and autonomous generative process. The latter processes usually produce more interesting and valuable ideas, but are also more noisy and prone to producing large amounts of low quality results. Having the ability to automatically filter out the low quality results or to rank the generated ideas would be highly beneficial.

The work[2] presented in this paper is a continuation of our studies [9] that address the problem of automated assessment of *What-if* ideas by machine learning. We are targeting three objectives: (I) to examine whether the problem of learning to differentiate among *What-ifs* that will be appreciated as good or bad is at all possible, then (II) to learn a classifier that could be used for this task in our setting, and (III) to gain some insight into what makes the fictional ideas to be assessed as good or bad.

[1] http://www.whim-project.eu/

[2] Extended results are reported also in Deliverable D6.3 of project WHIM.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Data

Data for the purposes of our analysis and modelling was obtained by crowdsourcing[3] the *What-if* labelling task to obtain human evaluations of the computer-generated *What-ifs*. For this purpose, we prepared an online questionnaire system that created questionnaires from a pool of *What-ifs* from various WHIM generators and collected the responses. The *What-ifs* in each questionnaire were selected at random from the given pool, but the system took care of dispersing the annotations among the *What-ifs* by favoring addition of items with fewer labels into each new questionnaire. Each questionnaire contained 15 *What-ifs*, which were shown in groups of 5. The system allowed completion of the crowdsourcing task only if respondents labelled all the *What-ifs* and provided textual comments for at least two of the *What-ifs* in each questionnaire. Each user of the platform could fill-out at most 50 questionnaires. In the crowdsourcing process 3,754 *What-ifs* were annotated, but in the work described in this paper, we use a subset (3,203) of all the annotated *What-ifs* for which we have available all the necessary features.

2.1 Annotations

The annotators scored the *What-ifs* on a 5-point Likert scale based on two criteria: (I) narrative potential/thought provoking, and (II) novelty/surprise. Each *What-if* was labeled between 4 and 8 times (mostly 7 times) by different annotators. The appropriate annotator agreement measure for our setting [1] is the intra-class correlation[4] that in our case equals 0.123, which is low. We merged all the scores of one *What-if* by calculating their median and average value. Finally, we also discretized the label values into two classes according to medians of scores ([1,2]: *low* and [4,5]: *high*) and obtained the final dataset of 1,298 items, out of which 591 are labelled as positive and 707 as negative (the majority baseline is 54.47%) regarding the narrative potential, which is the target label in the work presented in this paper.

Table 1 presents some of the best- and worst-scored *What-if* sentences together with their narrative potential (NP) and novelty/surprise (N/S) scores. The *What-ifs* are ordered by median score of the narrative potential. The second ordering criterion is the average score of the narrative potential.

Table 1: *What-ifs* ordered in terms of the median and average score of the narrative potential.

Rank	*What-if* sentence	NP Mdn	NP Avg	N/S Mdn	N/S Avg
1	What if there was a little cat who couldn't meow?	5	4.714	4.5	4.167
2	What if an astrologer observing a beautiful star becomes an eyewitness observing a horrid crime?	5	4.500	4	3.333
3	What if nonviolent hippies were to reject vegetarianism, take up their hooliganism and become violent hooligans?	5	4.333	4	4.143
...	...	...	...	...	...
3,201	What if there was an old fish, who couldn't swim anymore, which he used to do for relaxation, so decided instead to play frisbee?	1	1.333	1	2.333
3,202	What if a surveyor using a still tripod becomes a marketer using an animated mascot?	1	1.000	1	1.667
3,203	What if the princes that inherit grand thrones rule over the lowliest serfs?	1	1.000	1	1.333

2.2 Features

There are three kinds of data features that we generate for each *What-if*: (I) the *BoW*(Bag of words) features, which are essentially the frequencies of words, (II) the *linguistic* features: ambiguity, rhyming, length, sentiment[5] and the frequencies of adjectives and verbs as described in [9] and (III) the *narrative-based* features [6], which are experimental metrics of WHIM that aim to describe the narrative characteristics of an idea. The original numerical values of narrative-based and linguistic

[3]We used the services of the commercial platfrom CrowdFlower (https://www.crowdflower.com/).

[4]For items with ≥ 7 scores using the *irr* library for *R* and parameters: model="oneway", unit="average".

[5]Negative, neutral and positive sentiment values, calculated by employing the *TwoPlaneSVMbin* sentiment classifier as described in [5].

features (except sentiment) were transformed into the low and high values by using their median values as separators. We are interested particularly in the latter two types of more elaborate features and in Table 2 we provide some feature importance measures (correlation and Relief) for them.

Table 2: Correlation and Relief on discretized data for the 15 features with correlation above 0.05 and a positive Relief score.

(a) correlation		(b) Relief	
FictionalRatio	0.19290	FictionalRatio	0.14561
RealityDistorsionRatio	0.19290	RealityDistorsionRatio	0.14561
MainCharacterEventsRatio	0.12178	StoryCharacters	0.09306
ValenceAverage	0.11405	sentiment	0.09168
RichLife	0.11167	MainCharacterEventsRatio	0.04854
Handicap	0.10844	JointWordsProbability	0.02619
StoryCharacters	0.10521	JointWordsProbabilityMinimum	0.02542
ResolutionTriggerRatio	0.10017	ExplicitFact	0.01618
RatioCharacters	0.08225	Length	0.01618
ConflictTriggerRatio	0.07770	ConflictTriggerRatio	0.00847
ValenceSum	0.07496	RatioCharacters	0.00154
JointWordsProbabilityMinimum	0.07410	length	0.00000
TotalStoriesGenerated	0.07102	DivergencyMinimum	0.00000
ambiguity	0.06487	Originality	0.00000
numAdj	0.05657	OriginalityAccurate	0.00000

Both the correlation and the Relief scores of the features are very low. The narrative-based features seem to be more promising, as among the features that appear in both the top 10 for correlation and the top 10 for Relief, all are of this kind: *FictionalRatio*, *RealityDistortionRatio*, *MainCharacterEventsRatio*, *StoryCharacters* and *ConflictTriggerRatio*. Two narrative-based features stand out: *FictionalRatio* and *RealityDistortionRatio*, while from the linguistic features, the *ambiguity* is best scored by correlation and *sentiment* by Relief.

3 Models and evaluation

Based on the data described in Section 2 we constructed and evaluated human assessment models for *What-ifs* using the Support Vector Machine (SVM) algorithm [7][6] with various feature sets: using words, linguistic features and narrative-based features alone or in different combinations. As in our previous study [9], we also performed tokenization, stemming, extracted unigrams, removed features which appeared less then 2 times in the dataset, and applied the normalized term frequency (TF) weighting scheme.[7]

The performance evaluation was conducted by 10-fold cross validation. The results of settings which employ different types of features are shown in Table 3. Results indicate that it is highly beneficial to use words as features, while the differences in performance between the feature sets which combine words and various linguistic or narrative-based features are in most cases minor. The best performing feature set combines words and the *JointWordsProbability* feature. The last two rows in Table 3 present the results of a case in which we used also the (internal) information on the generator process of a *What-if*. Namely, as there are differences in score distributions of items from different processes, the information on the generation processes could crucially influence the classification performance. The results indicate that the origin of a *What-if* is informative, but it does not explain all of the information provided by the *BoW* features.

4 Conclusions

Modelling the human evaluation of *What-ifs* is a very challenging task. Our experiments indicate that the human *What-if* assessment could be modeled to some extent. Using all the available features, the best performing classifiers in our experiments were consistently above the baseline performance for

[6] We employed the binary SVM^{light} [2] implementation with linear kernel.

[7] The experiments were executed using the LATINO library (`https://github.com/LatinoLib/LATINO`).

Table 3: Results of 10-fold cross validation experiments using different feature sets. Results are presented in terms of average accuracy.

Features	Accuracy
words	65.04%
all linguistic + all narrative-based	62.16%
words + all linguistic + all narrative-based	64.03%
all narrative-based	62.32%
words + all narrative-based	63.59%
words + ConflictTriggerRatio	65.12%
words + Divergency	64.81%
words + DivergencyMinimum	64.96%
words + EndNames	64.89%
words + EndNamesRatio	64.89%
words + Evolution	64.97%
words + ExplicitFact	65.05%
words + FictionalAdditionsRatio	64.89%
words + FictionalRatio	65.04%
words + Handicap	64.89%
words + JointWordsProbability	65.19%
words + JointWordsProbabilityMinimum	64.96%
words + Length	65.05%
words + MainCharacterEventsRatio	64.96%
words + Originality	64.96%
words + OriginalityAccurate	64.81%
words + RatioCharacters	64.97%
words + RealityDistorsionRatio	65.04%
words + ResolutionTriggerRatio	64.96%
words + RichLife	64.89%
words + SettingQuality	64.89%
words + StoryCharacters	64.81%
words + TotalStoriesGenerated	64.89%
words + ValenceAverage	64.81%
words + ValenceSum	64.81%
all linguistic	55.88%
words + all linguistic	64.66%
words + length	64.96%
words + ambiguity	64.96%
words + rhyming	64.82%
words + number of adjectives	64.89%
words + number of verbs	64.74%
words + sentiment	64.96%
generator	63.04%
words + generator	64.96%

more than 10%, which means that the problem is not impossible to model and learn upon. However, the achieved performance scores were not high in absolute terms, so the effect of using the developed classifiers for automated *What-if* assessment might not be very notable in practice.

Assessment of classification performance, however, was not the only aim of our work. We have gained an insight into which features of *What-ifs* might be more important and which are probably irrelevant. The most important set of features are the words that are used in the *What-if* sentences. It seems that it might be beneficial to extend them also with some features from the much smaller set of the narrative-based ones, but the current results are too weak to draw definitive conclusions. As a lot of focus in WHIM was on development of the narrative-based features, it is encouraging to see that some of the experiments indicate that certain narrative-based features seem to be potentially useful for this kind of modelling.

Acknowledgments

This research was supported by the Slovene Research Agency through the research program Knowledge Technologies (Grant P2-0103) and through EC funding for the project WHIM 611560 by FP7, the ICT theme, and the Future Emerging Technologies FET programme.

References

[1] Kevin A Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012.

[2] Thorsten Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[3] Maria Teresa Llano, Simon Colton, Rose Hepworth, and Jeremy Gow. Automated fictional ideation via knowledge base manipulation. *Cognitive Computation*, 8(2):153–174, 2016.

[4] Maria Teresa Llano, Rose Hepworth, Simon Colton, Jeremy Gow, John Charnley, Nada Lavrač, Martin Žnidaršič, Matic Perovšek, Mark Granroth-Wilding, and Stephen Clark. Baseline methods for automated fictional ideation. In *Proceedings of the Fifth International Conference on Computational Creativity*, Ljubljana, Slovenia, jun 2014. Jožef Stefan Institute, Ljubljana, Slovenia, Jožef Stefan Institute, Ljubljana, Slovenia.

[5] Igor Mozetič, Miha Grčar, and Jasmina Smailović. Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE*, 11(5):e0155036, 2016.

[6] A. Tapscott, J. Goméz, Carlos León, J. Smailović, Martin Żnidaršič, and Pablo Gervás. Empirical evidence of the limits of automatic assessment of fictional ideation. In *Fifth International Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI 2016)*, 2016.

[7] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[8] Tony Veale. Coming good and breaking bad: Generating transformative character arcs for use in compelling stories. In *Proceedings of the Fifth International Conference on Computational Creativity*, Ljubljana, Slovenia, jun 2014. Josef Stefan Institute, Ljubljana, Slovenia, Josef Stefan Institute, Ljubljana, Slovenia.

[9] Martin Žnidaršič and Jasmina Smailović. Classification of fictional *What-if* ideas. In *Proceedings of the 18th International Multiconference Information Society – IS 2015*, Ljubljana, Slovenia, 2015.