Computational Exploration of the Linguistic Structures of Future-Oriented Expression: Classification and Categorization

Aiming Nie[1,2], Jason Shepard[2], Jinho Choi[1], Bridget Copley[3], Phillip Wolff[2]
[1]Dept. of Computer Science, Emory University, [2]Dept. of Psychology, Emory University
[3]Structures Formelles du Language, 4CNRS / Universite Paris 8, Paris, France 30322
{anie, jason.s.shepard, jinho.choi, pwolff}@emory.edu, bridget.copley@sfl.cnrs.fr

Abstract

English, like many languages, uses a wide variety of ways to talk about the future, which makes the automatic identification of future reference a challenge. In this research we extend Latent Dirichlet allocation (LDA) for use in the identification of future-referring sentences. Building off a set of hand-designed rules, we trained a ADAGRAD classifier to be able to automatically detect sentences referring to the future. Uni-bi-trigram and syntactic rule mixed feature was found to provide the highest accuracy. Latent Dirichlet Allocation (LDA) indicated the existence of four major categories of future orientation. Lastly, the results of these analyses were found to correlate with a range of behavioral measures, offering evidence in support of the psychological reality of the categories.

1 Introduction

Early formal work on tense such as (Prior, 1967) treated tenses as logical operators; this approach, however, could not correctly account for complex tenses, and was superseded by relational accounts (Reichenbach, 1947; Hornstein, 1990; Klein, 1997). However, these frameworks too fall short to the extent that they only posit three times (corresponding to the speech time, a reference time, and a time at which an event happens (Reichenbach's S, R, and T respectively). Natural language, however, can accommodate more than three times, as in *Before yesterday, Mary had been going to go to Paris on Friday*. In a Reichenbachian system, the reference time referred to by this sentence, would be *yesterday*, but then not only is there the event time of her going to Paris, but a time before yesterday is needed for Mary's plan as well. The future orientation (that is, the future relationship between reference time and event time) of such a sentence cannot be modeled in Reichenbach's system. Such examples indicate that a analysis with greater sensitivity to linguistic structure is needed if reference to the future is to be identified and modeled.

In this paper we use the syntactic properties of a sentence to identify references to the future. We also examine how references to the future might be diagnostic of a person's psychological wellbeing. In particular, we hypothesize that references to the future reflect, in part, a person's future-orientation, that is the proportion of time a person's thoughts concern the future.

Apparently, reference to future has sparked the interests of many Psychologists. Recent researches suggest that future-oriented thinking is linked to physical and mental health, academic achievement, increased social involvement, and lower distress (Kahana et al., 2005; Aspinwall, 2005; Simons et al., 2004).

While future-oriented thought appears to play a central role in cognition, it's identification in languages such as English is not easily accomplished. As pointed out earlier, the absence of explicit and necessary morphology for the encoding of future reference often makes distinguish references to the future or present difficult to determine.

The goal of this research is to develop procedures for the automated detection of references to the future, even in the context of a mix of verbs with differ-

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 168–173,
Denver, Colorado, June 1, 2015.

ent tenses. Such procedures will allow linguists and psychologists to more effectively mine text from social media to better extract chains and causation, as well as, potentially determine a person's or group's state of wellbeing. To the best of our knowledge, this is the first time that a project of this kind has been done in English, though similar research has been conducted in Japanese (Nakajima et al., 2014).

2 Related work

Document classification has been a long researched topic. Tools and algorithms have been developed to enable people to classify pre-labeled documents. The approach in this paper is single-label text classification using ADAGRAD (Duchi et al., 2011a).

Later on, we explored Latent Dirichlet Modeling (Blei et al., 2003) on the basis of induced subtrees, which are commonly used in data mining, but not frequently seen in Natural Language Processing. Frequent Subtree Mining is a common data mining topic. Related algorithms such as TreeMiner, FreeQT have been developed to find most frequent structure in a given tree bank (Chi et al., 2005).

Similar approaches have been explored in Moschitti (2006)'s work on using subtrees as features for Support Vector Machine. We did not use his approach because were were not interested in the similarity between tree structures, but rather in the linguistic regularities implicit in the text. For this reason, we chose to use Varro algorithm developed by Martens (2010), to exhaustively generate subtrees.

3 Data

We used data collected through Amazon Mechanical Turk (MTurk). Participants were asked to write down their mind wanderings as follows:

> *Please think back to the last time you were thinking about something other than what you were currently doing. Please share with us what you were thinking about. If you found yourself thinking about many different things, please share with us as many of these things that you can remember.*

In addition to writing down their mind wanderings, participants (N = 795) also answered a series of behavioral survey questions related to anxiety, health, happiness, life and financial satisfaction. The task resulted in a total of 2007 sentences. Table 1 describes the distribution of our data.

The sentences were rated by three human raters. For each sentence, raters indicated whether the expression referred to the future and their level of confidence of their decision.

	Sentence	**Subtree**	**Token**
Future	867	164,772	11,910
Not Future	1140	196,049	15,228

Table 1: Total number of sentences, subtrees and tokens

We used the Stanford factored parser (Klein and Manning, 2002) to parse sentences into constituency grammar tree representations. Tokens were generated by a uni-bi-trigram mixed model. Subtree structures were generated using the Varro algorithm (Martens, 2010) with threshold k = 1 to include lexicons. For the future corpus, 2,529,040 subtrees were processed while for the non-future corpus 2,792,875 were processed. A subset of the subtrees were selected as *words* for the LDA analysis, as described in Martens (2009).

4 Examples

While there are many cases of grammatical future marking (i.e., *will*, *be going to*) and lexical future meaning (e.g., *plan*, *want*, *need*, *tomorrow*, *goal*, *ambition*), many of the ways people use to refer to the future do not fall into one of these two types of linguistic categories.

For example, as we have seen, it's possible to have future reference without an obvious grammatical or lexical way of referring to the future. One way of doing this is with so-called *futurate* sentences (Copley, 2009; Kaufmann, 2005), such as *Mary is going to Paris*, which can refer to a contextually-provided future time (e.g., *tomorrow*). Another way to refer to the future without grammatical or lexial means is to use a *wh*-question word with an infinitive, such as in *I'm thinking about what to eat*. Such cases will be missed by ngram approaches.

Secondly, relying purely on lexical targets will not work well when sense disambiguation is required. Modals in English can have multiple meanings (Palmer, 1986):

I was thinking about the local news because they were showing what the weather would be like.

I was thinking about my life and marriage and how much money or lack of plays a role in my obligations, and what my husband would do if I died.

Both sentences have the modal word *would*. Many cases of *would* are "sequence-of-tense" *woulds*, as in the first sentence above. That is, they should really be seen as *will* in the past; the past-tense marking inherent to *would* is functioning as a kind of tense agreement with the main clause past. The future orientation provided by *would* is future with respect to the past reference time. However, the *would* in the second sentence is not a *will* of a past reference time, but picks out a "less-vivid" future relative to the present reference time (Iatridou, 2000).

5 Classification

5.1 Syntactic structural rules

We used the constituency grammar rules generated by Wolff and Copley. Rules were generated on the basis of linguistic theory, and then later refined on the basis of analyses of the false positives and misses.

The rules were instantiated in the Tregex pattern language (Levy and Andrew, 2006), which could then be used to find matching structures in the parsed sentences. There were 39 future-related rules, 16 past-related rules, and 3 present-related rules. The rules varied from the purely syntactic to the lexical, with a number of rules containing of mix of both. Syntactic information helped to disambiguate the senses of the modal verbs. Fourteen of the future-related rules emphasized the modal verbs. Rules are released online at `https://github.com/clir/time-perception`.

5.2 Adaptive sub-gradient descent

To build statistical models, we used a stochastic adaptive subgradient algorithm called ADAGRAD that uses per-coordinate learning rates to exploit rarely seen features while remaining scalable (Duchi et al., 2011b). This is suitable for NLP tasks where rarely seen features often play an important role and training data consists of a large number of instances with high dimensional features. We use the implementation of ADAGRAD in ClearNLP (Choi, 2013) using the hinge-loss, and the default hyperparameters (learning rate: $a = 0.01$, termination criterion: $r = 0.1$).

5.3 Experiments

Our experiment consists of four parts. First, we used the Tregex-based rule discussed in section 5.1 to determine whether the sentences referred to the future. Each sentence was matched against all rules, and an odd ratio score was calculated on the basis of the equation in (1).

$$\frac{Future}{Future + Past + Present} \quad (1)$$

We used this as our baseline classifier. In the second part of the experiment, we converted the rule matches into vector: matches were coded as 1's, absences as 0's.

In the third part of the experiment, we used a more traditional uni-bi-trigram mixed model as features for ADAGRAD. The extracted number of tokens from the corpus are represented in Table 1. Finally, we mixed the ngram features with rule-based features to train the final classifier. All classifiers were trained through a 5-fold cross-validation process. In the case of the human raters, we selected the label that was selected by 2 of the 3 raters. Table 3 shows the results of our classification.

	odd-ratio	**human**
accuracy	70.75	87.38[1]

Table 2: Simple Rule and Human Performance

6 Categorization

6.1 Induced subtree

Three types of subtrees are generally researched in subtree mining: bottom-up subtrees, induced subtrees, and embedded subtrees. They are ranked in order from the most restrictive to the most free

[1] Due to the fact that the corpus was slowly built over a year, and confidence rating task was later added to the rating task, thus only tested over 1034 sentences.

rules	ngram	ngram + rules
75.12	77.61	83.33
71.14	81.09	78.86
75.56	83.54	83.29
74.81	79.30	82.04
74.81	80.55	84.79
74.29	80.42	82.46

Table 3: 5-fold Cross-Validation: ADAGRAD Classifier Performance in Accuracy

form. Bottom-up subtree mining does not capture the transformations of a sentence, while embedded tree mining breaks a sentence structure down into units that are often unhelpful. Given these limitations, we used induced subtree mining, as recomended in (Martens, 2009).

After the initial extraction, we combined subtrees from the future, past, and present corpora to produce 322,691 subtrees. Each subtree's weights were calculated using the frequency of the subtree appearing in the future corpus divided by total number of sentence in future corpus minus the same subtree appearing in non-future corpora divided by total number of sentences in non-future corpus.

Linguists have long argued that syntactic constructions encode meaning (Grimshaw, 1990; Levin and Hovav, 1995). We argue that by using the subtree structures to represent a sentence, the components of meaning associated with a syntactic construction can be teased apart. The components of meaning associated with these subtrees can then be inferred using procedures such as latent dirichlet allocation (LDA).

6.2 Recursive LDA

We implemented a procedure called recursive LDA in which LDA was performed iteratively within new topics. One of the obstacles of modelling data using LDA is that the number of topics must be chosen in advance. Therefore it is very necessary to understand the properties of the data being modelled and choose a number of categories appropriately. Variations and extensions of LDA should also be modelled to reflect the characteristics of the space and the categories being modelled. With this in mind, we hypothesize that the total future-oriented reference space could be divided into a small number of categories and within each semantic category, future-oriented reference relate to each other will form more specific categories. In comparison to a similar extension: hLDA (Griffiths and Tenenbaum, 2004), rLDA provides better control to researchers, and is more suitable to discover categories on well-studied problems.

To run rLDA, we selected subtrees with weights larger than 0 (N = 21,156; 6.56% of the total generated subtree structures) as our features (words) and sentences identified as referring to the future as our collections (N = 867)(documents). Specifically, LDA was run on all of the subtrees with the goal of discovering 2 topics. The solution from this analysis was then used to divide the subtrees into two groups, and LDA was subsequently run again on each set of subtrees.

6.3 Experiments

We obtained 4 topics through two recursive run with LDA. All of which have significant statistical correlations with behavioral data. Two topics on the first level are labeled as topic A and topic B.

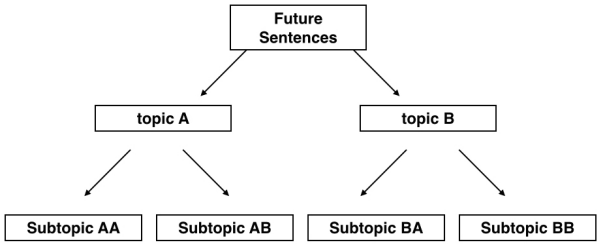

Figure 1: Recursive LDA Topic Hierarchy

The main semantic difference between A and B seemed to concern the distinction between open and fixed futures. Sentences in topic A indicate far fewer or more fixed choices, normally between just two choices. Sentences in topic B tend to include open-ended questions. Example sentences from these two sub-types are shown below:

Topic A - Fixed future:

I was thinking that I should not be playing Hay Day and I should do my work.

Last night I decided that I should travel to meet my aunt in Rhode Island as I haven't

	Topic AA	Topic AB	Topic BA	Topic BB
Age	.055	.397**	-.286**	-.167
Vividness	.157	.199	-.266*	-.100
Anxiety-State	.105	-.383**	.260	-.041
Anxiety-Trait	.050	-.342*	.247	-.008
Financial Satisfaction	.114	.326*	-.364**	-.032
Control over Life	.107	-.299**	.149	.039

Table 4: Correlation Table Between LDA Topics and Behavioral Data. Due to the iterative design of our survey, we did not have a complete behavioral question section till the end of our data collection. 146 people accounting for 18.36% of the total sample participated in the behavioral question research, and a subset of 81 people had future sentences in their response. Only content items that correlated with at least one category reported. $*p < .01$, $**p < .002$

seen her in a long time.

Topic B - Open Future:

At the same time I was thinking about what I was going to have for breakfast.

I was thinking about what I would cook for dinner tonight.

From the second level, more fine-grained topics emerged. Descending from topic A (fixed future), the two sub-types seemed to differ with respect to level of certainty: Topic AA tended to involve sentences conveying the notion of uncertainty, while Topic AB tended to involve sentences implying certainty. From Table 4 People, who construct future sentences with high certainty, have less control over life, scored lower on the trait and state anxiety inventory (Spielberger, 2010).

Topic AA - Uncertainty:

I was thinking about a trip that I may take at the end of the summer.

I was wondering if we would end up together and thinking about the fact that something that can seem so certain now may not be in the future.

Topic AB - Certainty:

I was making my wife 's lunch to take to work , and I was thinking about playing golf this weekend .

I am getting married in April , and there is a bunch of stuff left to be done .

Topic B appeared to be mostly about an open future. Its sub-types seemed to differ with respect to the notion of constraint: Topic BA seemed to consist of sentences about an unconstrained future while Topic BB seemed to concern sentences implying a constrained future. Our categorization matches with behavioral data in Table 4. People using unconstrained future sentence constructs rated their future as less vivid. They also were younger and had lower financial satisfaction.

Topic BA - Unconstrained:

I was thinking about what I should do for the rest of the day.

I was thinking about what I should animate for my next cartoon.

Topic BB - Constrained:

Two hours ago I was debating what I should have for lunch and what I should watch while I was eating.

I was thinking about a girl I would like to meet , what we would do , and how long we would do it.

7 Conclusion

In this research we leveraged recent developments in linguistic theory (Iatridou, 2000; Condoravdi, 2002; Copley and Martin, 2014) to build an automated system capable of discovering different ways of expressing the future. Specifically, we trained a ADAGRAD classifier to a relatively high level of accuracy and examined the number of topics associated with references to the future through the use of recursive

LDA. Finally, we established the psychological reality of our topics via comparisons to behavioral measures.

8 Acknowledgements

This research was supported by a grant from the U Penn / John Templeton Foundation to B. Copley and P. Wolff.

References

L. G. Aspinwall. 2005. The psychology of future-oriented thinking: From achievement to proactive coping, adaptation, and aging. *Motivation and Emotion*, 29(4):203–235.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Yun Chi, Richard R Muntz, Siegfried Nijssen, and Joost N Kok. 2005. Frequent subtree mining-an overview. *Fundamenta Informaticae*, 66(1):161–198.

Jinho D Choi. 2013. Clearnlp.

Cleo Condoravdi. 2002. Temporal interpretation of modals. In David Beaver, Stefan Kaufmann, Brady Clark, and Luis Casillas, editors, *Stanford Papers on Semantics*. CSLI Publications, Palo Alto.

Bridget Copley and Fabienne Martin, editors. 2014. *Causation in Grammatical Structures*. Oxford University Press.

Bridget Copley. 2009. *The semantics of the future*. Routledge.

John Duchi, Elad Hazan, and Yoram Singer. 2011a. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

John Duchi, Elad Hazan, and Yoram Singer. 2011b. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *The Journal of Machine Learning Research*, 12(39):2121–2159.

DMBTL Griffiths and MIJJB Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:17.

Jane Grimshaw. 1990. *Argument structure.* the MIT Press.

Norbert Hornstein. 1990. *As Time Goes By*. MIT Press.

Sabine Iatridou. 2000. The grammatical ingredients of counterfactuality. *LI*, 31:231–270.

E. Kahana, B. Kahana, and J. Zhang. 2005. Motivational antecedents of preventive proactivity in late life: Linking future orientation and exercise. *Motivation and emotion*, 29(4):438–459.

Stefan Kaufmann. 2005. Conditional truth and future reference. *Journal of Semantics*, 22(3):231–280, August.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10.

Wolfgang Klein. 1997. *Time in Language*. Routledge, New York.

Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface*, volume 26. MIT press.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234. Citeseer.

Scott Martens. 2009. Quantitative analysis of treebanks using frequent subtree mining methods. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 84–92. Association for Computational Linguistics.

Scott Martens. 2010. Varro: an algorithm and toolkit for regular structure discovery in treebanks. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 810–818. Association for Computational Linguistics.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *EACL*, volume 113, page 24.

Yoko Nakajima, Michal Ptaszynski, Hirotoshi Honma, and Fumito Masui. 2014. Investigation of future reference expressions in trend information. In *Proceedings of the 2014 AAAI Spring Symposium Series,Big data becomes personal: knowledge into meaning–For better health, wellness and well-being*, pages 31–38.

F. R. Palmer. 1986. *Mood and modality*. Cambridge University Press, Cambridge.

Arthur Prior. 1967. *Past, Present, and Future*. Oxford University Press, Oxford.

Hans Reichenbach. 1947. *The tenses of verbs*. na.

J. Simons, M. Vansteenkiste, W. Lens, and M. Lacante. 2004. Placing motivation and future time perspective theory in a temporal perspective. *Educational Psychology Review*, 16(2):121–139.

Charles D Spielberger. 2010. *State-Trait Anxiety Inventory*. Wiley Online Library.