

Extracting Human Goals from Weblogs

Mark Kröll
Graz University of Technology
Inffeldgasse 21a
8010 Graz, Austria
mkroell@tugraz.at

Markus Strohmaier
Graz University of Technology and Know-Center
Inffeldgasse 21a
8010 Graz, Austria
markus.strohmaier@tugraz.at

Abstract

Knowledge about human goals has been found to be an important kind of knowledge for a range of challenging problems, such as goal recognition from peoples' actions or reasoning about human goals. Necessary steps towards conducting such complex tasks involve (i) acquiring a broad range of human goals and (ii) making them accessible by structuring and storing them in a knowledge base. In this work, we focus on extracting goal knowledge from weblogs, a largely untapped resource that can be expected to contain a broad variety of human goals. We annotate a small sample of weblogs and devise a set of simple lexico-syntactic patterns that indicate the presence of human goals. We then evaluate the quality of our patterns by conducting a human subject study. Resulting precision values favor patterns that are not merely based on part-of-speech tags. In future steps, we intend to improve these preliminary patterns based on our observations.

1 Knowledge about Human Goals

Knowledge about human goals has been found to be an important kind of knowledge for a range of challenging research problems, such as goal recognition from people's actions, reasoning about people's goals or the generation of action sequences that implement goals (planning) [Schank and Abelson, 1977]. In contrast to other kinds of knowledge, e.g. commonsense, knowledge about human goals provides a different perspective on textual resources putting more emphasis on future aspects and activities. We regard the acquisition of this knowledge as a first step towards conducting complex tasks such as planning.

Regardless whether the knowledge to extract is about human goals, commonsense [Liu and Singh, 2004] or the world in general [Schubert and Tong, 2003; Clarke, 2009], the acquisition process often includes the application of indication and extraction patterns. Moreover, knowledge acquisition approaches differ in how much manual intervention is necessary (or desired) in the knowledge acquisition process. Existing approaches include utilizing human knowledge engineering [Lenat, 1995], volunteer-based [Liu and Singh, 2004], game-based [Lieberman et al., 2007; von Ahn, 2006] or semiautomatic approaches [Eslick, 2006]. Yet, in this paper we are interested in approaching the question how knowledge about human goals can be automatically derived from social media text, in our case weblogs. To give an example, here

is a snippet of a blog post where human goals are underlined:

Last September, we moved into our new home. I had plans for this home, the first house--not apartment--my husband and I would live in. I was going to refinish some hand-me-down furniture we have, and I was going to plant a wonderful garden, starting with bulbs that would bloom in the spring. Crocuses, hyacinths, tulips--all of my favorites. And I would know, all winter long, that they were sleeping in the dark, cold soil, waiting to awake with the first light and warmth of spring

Though weblogs exhibit some disadvantages when it comes to quality issues, e.g., textual content is prone to noise, we can expect that weblogs contain a broad variety of human goals. In the remaining part, we describe our approach to address goal extraction from open text by deriving and evaluating a first set of lexico-syntactic patterns. We then discuss strengths and weaknesses of our patterns based on a small human subject study in order to improve them in future steps.

2 Patterns To Extract Human Goals

We employ and adapt the definition from [Tatu, 2005] who defines human goals as: "*Expressions of a particular action that shall take place in the future, in which the speaker is some sort of agent.*" The following, exemplary sentence taken from the blog post snippet presented above: "I was going to refinish some hand-me-down furniture" indicates the person's intention to prettify some furniture. In contrast to Tatu's definition, we do not include information about the speaker into our patterns to keep them simple. However, the idea to include this kind of information is discussed in Section 3.3. When comparing our setup to [Tatu, 2005]'s, we can observe three differences. Firstly, the author developed part-of-speech patterns by annotating and examining samples from the Brown corpus. Working on the Brown corpus is advantageous because this corpus has already been tagged – the chance of getting incorrect part-of-speech tags is thereby reduced. Secondly, the language used in the Brown corpus is different than language used in weblogs. Thirdly, [Tatu, 2005]'s motivation to address challenges in question answering (QA). She expected that sentences containing expressions of human goals are better suited to answer a certain kind of questions. Textual resources in the QA domain exhibit other characteristics than weblogs, for instance, people use weblogs to tell stories or write diary-like entries. We hypothesize that extraction patterns yield

different results depending on weblog characteristics, e.g., does the weblog contain a story-like structure or not?

We followed a common path to acquire knowledge from textual resources by manually examining the textual environment to identify appropriate patterns [Hearst, 1992]. As a first step, we drew a small, random sample (~100 blog posts) from the ICWSM 2009 Spinn3r Dataset [Burton et al., 2009] and annotated the textual contents according to the above definition. The annotation task was conducted by one of the authors and an undergraduate student. Table 1 illustrates ten resulting, lexico-syntactic patterns based on these annotations which are partly inspired by patterns by [Tatu, 2005]. She employs these patterns to identify sentences containing intentional expressions in order to build up a training set for further experiments. Part-of-speech tags throughout this paper are consistent with the Penn Treebank Tag Set.

Table 1: Lexico-syntactic patterns to identify and extract human goals and matching instances. (*) denotes no, one or several occurrences, (+) denotes at least one occurrence, (?) denotes one optional occurrence and (|) denotes a logical OR.

Nr.	Lexico-Syntactic Patterns	Matching Instances
1	<VB VBZ> <TO> <VB>	needs/VBZ to/TO organize/VB
2	<NN.*> <TO> <VB>	alcohol/NN to/TO get/VB
3	<JJ> <TO> <VB.*>	available/JJ to/TO read/VB
4	<VB> <DT> <NN.*>	find/VB a/DT keyboard/NN
5	<WANT> <TO> <VB>	wanted/VBD to/TO kill/VB
6	<INTEND> <TO> <VB>	intend/VBP to/TO quit/VB
7	<INTENT PURPOSE GOAL OBJECTIVE><VBZ><TO><VB NN.*>*	goal/NN is/VBZ to/TO eat/VB
8	<LIKE> <TO> <VB.*>	like/VB to/TO share/VB
9	<WANT> <PRP> <TO> <VB>	wants/VBZ them/PRP to/TO go/VB
10	<GET> <PRP> <DT>? <NN.*> <VB.*>	get/VB you/PRP to/TO purchase/VB

In the next section, we apply our extraction patterns to a larger sample of weblogs. We then evaluate the quality of every pattern by calculating precision values.

3 Quality & Characteristics

In this section, we briefly describe our data preparation steps and pattern matching process. We report precision results of preliminary study on a set of ~205.000 blog posts and discuss observed weaknesses of our patterns. We conclude this section with suggesting several possibilities to improve and extend the patterns to extract knowledge about human goals.

3.1 Data Sets

For our experiments, we used the ICWSM 2009 Spinn3r Dataset which comprises 44 million blog posts made between August 1st and October 1st, 2008. We randomly drew ~205.000 blog posts and further separated them into two datasets – one with posts containing stories – one with posts containing non-stories. We hypothesize that blog posts telling a story contain more human goals than other blog posts. We use work from [Gordon and Reid, 2009] that defines a story as a series of causally related events in the past. They developed an automatic algorithm to identify blog posts most likely containing a story (reported precision values up to 75%). Moreover, they provide an index of all blog posts in the ICWSM 2009 Spinn3r Dataset that were classified as containing story-like structures. Using this information, we obtained two

datasets – one containing posts with stories (~3000) and one containing posts without stories (~202.000).

3.2 Data Preparation

We first extracted the content of the <description> field in the corresponding xml files of the random sample. Since the textual content of the weblogs was often messy, we had to clean it as preparation for the subsequent part-of-speech tagging. The cleaning procedure included removing html snippets and special characters. For the process of part-of-speech tagging and pattern matching, we used functionality of the Natural Language Processing Toolkit (NLTK¹) in combination with Python as programming language.

3.3 Strengths and Weaknesses of our Goal Extraction Patterns

We applied our patterns from Table 1 to two datasets (see Section 3.1) which were randomly drawn from the ICWSM 2009 Spinn3r Dataset (tiergroups 1-3).

Table 2 shows the number of matches per extraction pattern. The frequency numbers corroborate our hypothesis that there is a higher potential for the presence of human goals weblogs containing a story. Since there are ~67 times more blog posts containing non-stories than stories, the numbers are not directly comparable. In order to compare them, we calculate the ratio of (number of found goal instances) vs. (number of blog posts). We notice that the ratio is always highly in favor of the blog posts containing stories. Consider for example ratios for the first pattern <VB|VBZ> <TO> <VB>: $486/3,000 = 0.16$ for stories vs. $6,220/202,000 = 0.03$ for non-stories.

Table 2 illustrates the number of matched goal instances per extraction pattern as well as precision values (sample size of 20) for both story and non-story content.

Lexico-Syntactic Patterns	Story Set (#3.000)		Non-Story Set (#202.000)	
	Freq.	Prec.	Freq.	Prec.
<VB VBZ> <TO> <VB>	486	0.1	6220	0
<NN.*> <TO> <VB>	2018	0	6661	0
<JJ> <TO> <VB.*>	677	0.05	5424	0.06
<VB> <DT> <NN.*>	1405	0.06	15129	0
<WANT> <TO> <VB>	398	0.53	3614	0.37
<INTEND> <TO> <VB>	10	0.6	86	0.5
<INTENT PURPOSE GOAL OBJECTIVE><VBZ><TO><VB NN.*>*	2	0.5	39	0.82
<LIKE> <TO> <VB.*>	36	0.16	592	0.18
<WANT> <PRP> <TO> <VB>	30	0.83	291	0.11
<GET> <PRP> <DT>? <NN.*> <VB.*>	16	0.25	47	0.32

For every pattern, an undergraduate student rated a maximum number of 40 matched instances whether a human goal is expressed or not. The student took the context (sentence boundary) into account when he rated the matched instances. The precision values for every pattern are calculated based upon 20 instances from story content and 20 instances from non-story content. In five cases, where the pattern matched fewer than 20 instances, the precision values are based on a slightly lower number of rated samples.

¹ <http://www.nltk.org/>

To discuss strengths and weaknesses, we group our patterns into three categories and provide positively and negatively rated instances per pattern category, i.e. true positives and false positives. The first category (Nr. 1 to 4) contains pure part-of-speech patterns, the second category (Nr. 5 to 8) contains part-of-speech patterns combined with goal keywords and patterns of the third group (Nr. 9 to 10) can be expected to extract not only goal knowledge but to extract additional information on the participants involved.

We can observe that precision values in the first category are low. Though these pure part-of-speech patterns produce a lot of matches, the matched instances appear too general and are therefore inappropriate to extract human goals. Moreover, positive examples are partly matched by other categories such as “want him to learn to ride a bike” which actually serves as positive examples for patterns Nr. 1 and Nr. 4. Table 3 shows true and false positives extracted by these patterns.

Table 3 shows true and false positives of extracted human goals (from patterns Nr.1-4).

Matched Instance	Context	Goal
learn/VB to/TO ride/VB	that want him to learn to ride a bike.	yes
have/VB to/TO agree/VB	I might have to agree on some levels	no
car/NN to/TO go/VB	We got in the car to go to the hospital	no
things/NNS to/TO load/VB	I just have a few more things to load	no
willing/JJ to/TO believe/VB	I am willing to believe in love	yes
ready/JJ to/TO take/VB	I was ready to take that chance with you	no
ride/VB a/DT bike/NNP	that want him to learn to ride a bike	yes
take/VB another/DT night/NN	I can't take another night of this	no

Patterns in the second category almost all achieved a precision value higher than 50% except for two patterns. The first one is pattern Nr. 8. When reviewing the matched instances, we find sentences such as: “She swims a lot and likes to drink lake water.” (see Table 4) where a person’s preferences are expressed. We would rather like to match sentences such as “I like to play soccer in the evening” implying an action that takes place in the future. Therefore, in order to improve this pattern, we could require the presence of certain temporal expressions such as ‘today’ or ‘tomorrow’. The second exception is pattern Nr. 7 (story content) where the moderate precision value is most likely due to the low number of matches. In case of the non-story content, this pattern yields high precision values demonstrating its usefulness.

Table 4 shows true and false positives of extracted human goals (from patterns Nr.5-8).

Matched Instance	Context	Goal
wanted/VBD to/TO go/VB	I never wanted to go back to school	yes
wants/VBZ to/TO do/VB	he wants to do it	no
intend/VBP to/TO get/VB	I intend to get up at 7:30	yes
intend/VB to/TO stay/VB	Jean, do you intend to stay here until morning?	no
goal/NN is/VBZ to/TO eat/VB	the ultimate goal is to eat the cookie	yes
goal/NN is/VBZ to/TO	on what makes a safe tire. With all your communications, you goal is to.	no
like/VB to/TO move/VB	We would like to move into this house sometime before the year 2020	yes
like/VBP to/TO do/VB	She swims a lot and likes to drink lake water.	no

Examining the remaining three negative examples in Table 4, we can gain further insights on how to improve the extraction patterns. The first negative instance: “He wants to do it” can be avoided by simply requiring the pattern to end in a verb phrase.

<WANT><TO><VB> → <WANT><TO>(<VB><DT>?<JJ>*<NN.>*)
 <INTEND><TO><VB> → <INTEND><TO>(<VB><DT>?<JJ>*<NN.>*)

The second negative instance: “Jean, do you intend to stay here until morning?” suggests to check whether the sentence is interrogative or not. A simple approach would be to take punctuation information into account, yet in case of weblog content punctuations might not always be provided.

The third negative instance: “With all your communications, you goal is to” can be easily avoided by adapting the pattern to require at least one verb or noun after the part-of-speech tag <TO>. In the same step, one might think of including plural forms of the keywords “goal, purpose, intent and objective” into the pattern.

<GOAL><VBZ><TO><VB|NN.>* → <GOAL><VBZ><TO><VB|NN.>+

In summary, we can speculate that the more patterns take advantage of context information the more accurate the extracted instances are. True and false positives of patterns Nr.5-8 are illustrated in Table 4.

Patterns in the third category exhibit an additional characteristic compared to the other two categories. To give an example where pattern Nr.9 matched following chunk (wanted/VBD me/PRP to/TO buy/VB) that was part of the sentence: “he wanted me to buy him a new chair”. Besides the purchase of a chair as a future action, we learn something about the relation among the participants. This information can be used when trying to identify the goal carrier, i.e. the person who actually issues her goal.

Table 5 illustrates true and false positives that were extracted by patterns Nr. 9 and Nr. 10.

Table 5 shows true and false positives of extracted human goals (from patterns Nr. 9-10).

Matched Instance	Context	Goal
wants/VBZ me/PRP to/TO send/VB	He wants me to send him a hard copy	yes
wanted/VBD it/PRP to/TO be/VB	I wanted it to be about me.	no
get/VB him/PRP to/TO email/VB	Mum went to contact her former pupil and get him to email me	yes
get/VB you/PRP to/TO do/VB	really have better things to do, so I'm going to get you to do it	no

3.4 Potential Extensions

In the previous subsection, we described various improvement strategies that were based on examining a small sample of false positives. In this subsection, we discuss potential benefits of including other feature types than only keywords and part-of-speech tags.

At present state, many false positives are due to incorrect part-of-speech tagging. We intend to examine whether we could become independent from tagging quality by, for example, only using lexical and punctuation features. A

conceivable approach could be to (i) identify indicators such as “intend to” and (ii) take the remaining tokens till the next punctuation is reached. Regular expressions represent a means to implement this approach which is then to be evaluated.

A more complex approach involves identifying the verb’s agent to ensure the identified goal belongs to a person [Tatu, 2005]. Including this feature could avoid false positives such as: “The dog is going to bite the postman”. However, the annotation of open text with linguistic features such as semantic roles and predicate argument structures is challenging. Challenges include (i) incorrect part-of-speech tagging and (ii) lack of adequate tools to annotate semantic roles. Thus, we suggest employing linguistic resources such as PropositionBank [Palmer et al., 2005] and FrameNet [Baker et al., 1998] that might help to evolve our patterns. The main advantage, for example, of the PropositionBank corpus is that it is already annotated not only with part-of-speech tags and parse tree information but with predicate-argument structures as well. An interesting next step would be to apply our patterns to the PropositionBank corpus to learn how the additional, linguistic information can be included in our patterns.

4 Summing Up

In this paper, we present work in progress towards generating a knowledge base that contains a broad spectrum of human goals. By analyzing such a knowledge base of human goals, we could learn something about people’s current (crawling date) motivations and intentions. We expect that weblogs are an appropriate source for acquiring this knowledge. As a first step, we evaluate a set of simple patterns to identify human goals in weblogs. We intend to extend these patterns based on our observations, above all to include verb phrases – apparently the predominant carrier of human goal expressions. With regard to our objective of generating a knowledge base of human goals, we deem precision of our patterns more important than recall.

We reckon that aspects of this work could inform social applications e.g. search in weblogs. While traditional systems search and rank weblogs based on content information, information about the goal a blogger pursues has not yet been taken into account. One could imagine a blog search system that ranks weblogs based on shared goals. Knowledge about goals could also provide a novel way to identify weblog communities. Participants of these communities would not only share interests but also common goals.

Acknowledgments

Thanks to Daniel Lamprecht for participating in the data cleaning and annotation tasks. This work is funded by the FWF Grant P20269 *TransAgere* and the Know-Center.

References

[Baker et al., 1998] C. Baker, C. Fillmore and J. Lowe. The Berkeley FrameNet Project, in 'Proceedings of the 17th international conference on Computational linguistics', Association for Computational Linguistics, Morristown, NJ, USA, pp. 86–90, 1998.

[Burton et al., 2009] K. Burton, A. Java, and I. Soboroff. The ICWSM 2009 Spinn3r Dataset. In Proceedings of

the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA, May 2009.

[Clarke and Harrison, 2009] P. Clark and P. Harrison. Large-Scale Extraction and Use of Knowledge From Text. In The Fifth International Conference on Knowledge Capture, 2009.

[Eslick, 2006] I. Eslick. Searching for commonsense, Master’s thesis, Massachusetts Institute of Technology, 2006.

[Gordon and Swanson, 2009] A. Gordon and R. Swanson. Identifying Personal Stories in Millions of Weblog Entries. Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA, May 20, 2009.

[Hearst, 1992] M. Hearst. Automatic acquisition of hyponyms from large text corpora, in 'Proceedings of the 14th conference on Computational linguistics', Association for Computational Linguistics, Morristown, NJ, USA, pp. 539–545, 1992.

[Lenat, 1995] D. Lenat. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, (38)11:33-38, 1995.

[Lieberman et al, 2007] H. Lieberman, D. Smith, A. Teeters. Common Consensus: a web-based game for collecting commonsense goals. In Proceedings of the Workshop on Common Sense and Intelligent User Interfaces held in conjunction with the 2007 International Conference on Intelligent User Interfaces, IUI, 2007.

[Liu and Singh, 2004] H. Liu and P. Singh. ConceptNet - A practical commonsense reasoning tool-kit. BT Technology Journal, (22)4:211-226, 2004.

[Palmer et al., 2005] M. Palmer, D. Gildea and P. Kingsbury. The Proposition Bank: A Corpus Annotated with Semantic Roles, Computational Linguistics Journal, 2005.

[Schank and Abelson, 1977] R. Schank and R. Abelson R: Scripts, plans, goals, and understanding: an inquiry into human knowledge structures. Lawrence Erlbaum Associates, 1977.

[Schubert and Tong, 2003] L. Schubert and M. Tong. Extracting and evaluating general world knowledge from the Brown corpus, in 'Proceedings of the HLT-NAACL 2003 workshop on Text meaning', Association for Computational Linguistics, Morristown, NJ, USA, pp. 7–13, 2003.

[Singh et al., 2002] P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins and W. Zhu. Open Mind Common Sense: Knowledge acquisition from the general public. In Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, 1223--1237, Springer-Verlag London, UK, 2002.

[Tatu, 2005] M. Tatu. Automatic Discovery of Goals in Text and its Application to Question Answering. In 43rd Annual Meeting of the Association for Computational Linguistics, 'ACL 05', 2005.

[von Ahn, 2006] L. von Ahn. 'Games with a Purpose', Computer 39(6), 92–94, 2006.