Constructing An Anaphorically Annotated Corpus With Non-Experts: Assessing The Quality Of Collaborative Annotations.

Jon Chamberlain
University of Essex
School of Computer Science
and Electronic Engineering
jchamb@essex.ac.uk

Udo Kruschwitz
University of Essex
School of Computer Science
and Electronic Engineering
udo@essex.ac.uk

Massimo Poesio
University of Essex
School of Computer Science
and Electronic Engineering
poesio@essex.ac.uk

Abstract

This paper reports on the ongoing work of *Phrase Detectives*, an attempt to create a very large anaphorically annotated text corpus. Annotated corpora of the size needed for modern computational linguistics research cannot be created by small groups of hand-annotators however the ESP game and similar *games with a purpose* have demonstrated how it might be possible to do this through Web collaboration. We show that this approach could be used to create large, high-quality natural language resources.

1 Introduction

The statistical revolution in natural language processing (NLP) has resulted in the first NLP systems and components really usable on a large scale, from part-of-speech (POS) taggers to parsers (Jurafsky and Martin, 2008). But it has also raised the problem of creating the large amounts of annotated linguistic data needed for training and evaluating such systems.

This requires trained annotators, which is prohibitively expensive both financially and in terms of person-hours (given the number of trained annotators available) on the scale required.

Recently, however, Web collaboration has started to emerge as a viable alternative. Wikipedia and similar initiatives have shown that a surprising number of individuals are willing to help with resource creation and scientific experiments. The goal of the ANAWIKI project[1] is to experiment with Web collaboration as a solution to the problem of creating large-scale linguistically annotated corpora. We do this by developing tools through which members of our scientific community can participate in corpus creation and by engaging non-expert volunteers with a game-like interface. In this paper we present ongoing work on *Phrase Detectives*[2], a game designed to collect judgments about anaphoric annotations, and we report a first analysis of annotation quality in the game.

2 Related Work

Large-scale annotation of low-level linguistic information (part-of-speech tags) began with the Brown Corpus, in which very low-tech and time consuming methods were used. For the creation of the British National Corpus (BNC), the first 100M-word linguistically annotated corpus, a faster methodology was developed using preliminary annotation with automatic methods followed by partial hand-correction (Burnard, 2000).

Medium and large-scale semantic annotation projects (for wordsense or coreference) are a recent innovation in Computational Linguistics. The semi-automatic annotation methodology cannot yet be used for this type of annotation, as the quality of, for instance, coreference resolvers is not yet high enough on general text. Nevertheless the semantic annotation methodology has made great progress with the development, on the one end, of effective quality control methods (Hovy et al., 2006) and on the other, of sophisticated annotation tools such as Serengeti (Stührenberg et al., 2007).

These developments have made it possible to move from the small-scale semantic annotation projects, the aim of which was to create resources of around 100K words in size (Poesio, 2004b), to the efforts made as part of US initiatives such as Automatic Context Extraction (ACE), Translingual Information Detection, Extraction and Summarization (TIDES), and GALE to create 1 million word corpora. Such techniques could not be expected to annotate data on the scale of the BNC.

[1] http://www.anawiki.org

[2] http://www.phrasedetectives.org

Proceedings of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009, pages 57–62,
Suntec, Singapore, 7 August 2009.

2.1 Collaborative Resource Creation

Collaborative resource creation on the Web offers a different solution to this problem. The motivation for this is the observation that a group of individuals can contribute to a collective solution, which has a better performance and is more robust than an individual's solution as demonstrated in simulations of collective behaviours in self-organizing systems (Johnson et al., 1998).

Wikipedia is perhaps the best example of collaborative resource creation, but it is not an isolated case. The gaming approach to data collection, termed *games with a purpose*, has received increased attention since the success of the ESP game (von Ahn, 2006).

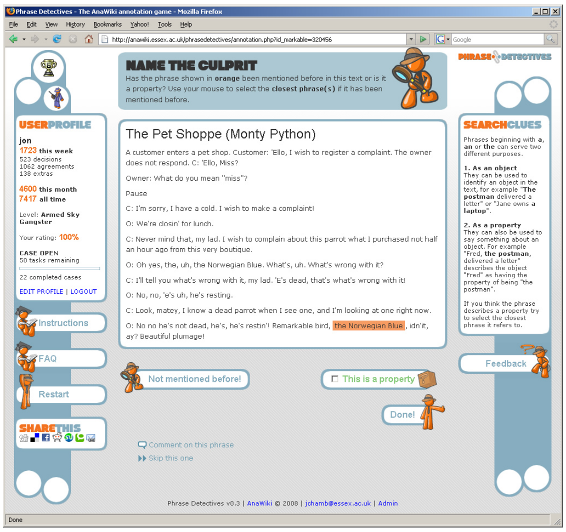
2.2 Human Computation

Human computation, as a more general concept than *games with a purpose*, has become popular in numerous research areas. The underlying assumption of learning from a vast user population has been largely the same in each approach. Users are engaged in different ways to achieve objectives such as:

- Assigning labels to items
- Learning to rank
- Acquiring structured knowledge

An example of the first category is the ESP game which was a project to label images with tags through a competitive game. 13,500 users played the game, creating 1.3M labels in 3 months (von Ahn, 2006). Other examples of assigning lables to items include Phetch and Peekaboom (von Ahn et al., 2006).

Learning to rank is a very different objective. For example user judgements are collected in the *Picture This* game (Bennett et al., 2009). This is a two player game where the user has to select the best matching image for a given query from a small set of potential candidates. The aim is to learn a preference ranking from the user votes to predict the preference of future users. Several methods for modeling the collected preferences confirmed the assumption that a consensus ranking from one set of users can be used to model another.



NAME THE CULPRIT

Has the phrase shown in **orange** been mentioned before in this text or is it a property? Use your mouse to select the **closest phrase(s)** if it has been mentioned before.

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

O: We're closin' for lunch.

C: Never mind that, my lad. I wish to complain about this parrot what I purchased not half an hour ago from this very boutique.

O: Oh yes, the, uh, the Norwegian Blue. What's, uh. What's wrong with it?

C: I'll tell you what's wrong with it, my lad. 'E's dead, that's what's wrong with it!

O: No, no, 'e's uh, he's resting.

C: Look, matey, I know a dead parrot when I see one, and I'm looking at one right now.

O: No no he's not dead, he's, he's restin'! Remarkable bird, the Norwegian Blue, idn'it, ay? Beautiful plumage!

SEARCHCLUES

Phrases beginning with **a**, **an** or **the** can serve two different purposes.

1. As an object
They can be used to identify an object in the text, for example "**The postman** delivered a letter" or "Jane owns **a laptop**".

2. As a property
They can also be used to say something about an object. For example "Fred, **the postman**, delivered a letter" describes the object "Fred" as having the property of being "the postman".

If you think the phrase describes a property try to select the closest phrase it refers to.

Phrase Detectives v0.3 | AnaWiki © 2008 | jchamb@essex.ac.uk | Admin

Figure 1: A screenshot of the Annotation Mode.

Phrase Detectives is in the third category, i.e. it aims to acquire structured knowledge, ultimately leading to a linguistically annotated corpus. Another example of aiming to acquire large amounts of structured knowledge is the Open Mind Commonsense project, a project to mine commonsense knowledge to which 14,500 participants contributed nearly 700,000 sentences (Singh, 2002).

Current efforts in attempting to acquire large-scale world knowledge from Web users include Freebase[3] and True Knowledge[4]. A slightly different approach to the creation of commonsense knowledge has been pursued in the Semantic MediaWiki project (Krötzsch et al., 2007), an effort to develop a 'Wikipedia way to the Semantic Web': i.e., to make Wikipedia more useful and to support improved search of web pages via semantic annotation.

3 The Phrase Detectives game

Phrase Detectives offers a simple graphical user interface for non-expert users to learn how to annotate text and to make annotation decisions (Chamberlain et al., 2008).

In order to use Web collaboration to create annotated data, a number of issues have to be addressed. First among these is motivation. For anybody other than a few truly dedicated people, annotation is a very boring task. This is where the promise of the game approach lies. Provided that a suitably entertaining format can be found, it may be possible to get people to tag quite a lot of data without them even realizing it.

[3] http://www.freebase.com/
[4] http://www.trueknowledge.com/

The second issue is being able to recruit sufficient numbers of useful players to make the results robust. Both of these issues have been addressed in the incentive structures of Phrase Detectives (Chamberlain et al., 2009).

Other problems still remain, most important of which is to ensure the *quality* of the annotated data. We have identified four aspects that need to be addressed to control annotation quality:

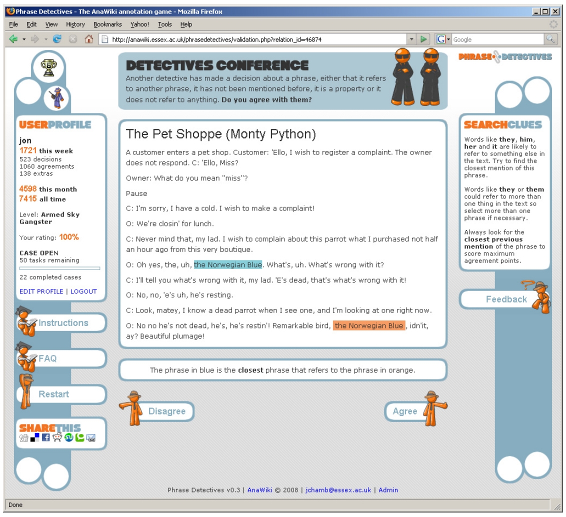
- Ensuring users understand the task
- Attention slips
- Malicious behaviour
- Genuine ambiguity of data

These issues have been addressed at the design stage of the project (Kruschwitz et al., 2009).

The goal of the game is to identify relationships between words and phrases in a short text. An example of a task would be to highlight an anaphor-antecedent relation between the markables (sections of text) *'This parrot'* and *'He'* in *'This parrot is no more! He has ceased to be!'* Markables are identified in the text by automatic pre-processing. There are two ways to annotate within the game: by selecting a markable that corefers to another one (Annotation Mode); or by validating a decision previously submitted by another player (Validation Mode).

Annotation Mode (see Figure 1) is the simplest way of collecting judgments. The player has to locate the closest antecedent markable of an anaphor markable, i.e. an earlier mention of the object. By moving the cursor over the text, markables are revealed in a bordered box. To select it the player clicks on the bordered box and the markable becomes highlighted. They can repeat this process if there is more than one antecedent markable (e.g. for plural anaphors such as *'they'*). They submit the annotation by clicking the *Done!* button. The player can also indicate that the highlighted markable has not been mentioned before (i.e. it is not anaphoric), that it is non-referring (for example, *'it'* in *'Yeah, well it's not easy to pad these Python files out to 150 lines, you know.'*) or that it is the property of another markable (for example, *'a lumberjack'* being a property of *'I'* in *'I wanted to be a lumberjack!'*).

In Validation Mode (see Figure 2) the player is presented with an annotation from a previous player. The anaphor markable is shown with the antecedent markable(s) that the previous player chose. The player has to decide if he agrees with this annotation. If not he is shown the Annotation Mode to enter a new annotation.


Phrase Detectives - The AnaWiki annotation game - Mozilla Firefox
File Edit View History Bookmarks Yahoo! Tools Help
http://anawiki.essex.ac.uk/phrasedetectives/validation.php?relation_id=46874
Google
PHRASE DETECTIVES
DETECTIVES CONFERENCE
USERPROFILE
SEARCHCLUES
Feedback
Instructions
FAQ
Restart
SHARETHIS
Disagree
Agree
Done

Another detective has made a decision about a phrase, either that it refers to another phrase, it has not been mentioned before, it is a property or it does not refer to anything. **Do you agree with them?**

jon
1721 **this week**
523 decisions
1060 agreements
138 extras

4598 **this month**
7415 **all time**

Level: **Armed Sky Gangster**

Your rating: 100%

CASE OPEN
50 tasks remaining

22 completed cases

EDIT PROFILE | LOGOUT

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

O: We're closin' for lunch.

C: Never mind that, my lad. I wish to complain about this parrot what I purchased not half an hour ago from this very boutique.

O: Oh yes, the, uh, the Norwegian Blue. What's, uh. What's wrong with it?

C: I'll tell you what's wrong with it, my lad. 'E's dead, that's what's wrong with it!

O: No, no, 'e's uh, he's resting.

C: Look, matey, I know a dead parrot when I see one, and I'm looking at one right now.

O: No no he's not dead, he's, he's restin'! Remarkable bird, the Norwegian Blue, idn'it, ay? Beautiful plumage!

The phrase in blue is the **closest** phrase that refers to the phrase in orange.

Words like **they**, **him**, **her** and **it** are likely to refer to something else in the text. Try to find the closest mention of this phrase.

Words like **they** or **them** could refer to more than one thing in the text so select more than one phrase if necessary.

Always look for the **closest previous mention** of the phrase to score maximum agreement points.

Phrase Detectives v0.3 | AnaWiki © 2008 | jchamb@essex.ac.uk | Admin

Figure 2: A screenshot of the Validation Mode.

In the game groups of players work on the same task over a period of time as this is likely to lead to a collectively intelligent decision (Surowiecki, 2005). An initial group of players are asked to annotate a markable. If all the players agree with each other then the markable is considered complete.

However it is likely that the first group of players will not agree with each other (62% of markables are given more than one relationship). In this case each unique relationship for the markable is validated by another group of players. This type of validation has also been proposed elsewhere, e.g. (Krause and Aras, 2009).

When the users register they begin with the training phase of the game. Their answers are compared with Gold Standard texts to give them feedback on their decisions and to get a user rating, which is used to determine whether they need more training. Contextual instructions are also available during the game.

The corpus used in the game is created from short texts including, for example, Wikipedia articles selected from the 'Featured Articles' and the page of 'Unusual Articles'; stories from Project Gutenberg including Aesop's Fables, Sherlock Holmes and Grimm's Fairy Tales; and dialogue texts from Textfile.com.

	Expert 1 vs. Expert 2	Expert 1 vs. Game	Expert 2 vs. Game
Overall agreement	94.1%	84.5%	83.9%
DN agreement	93.9%	96.0%	93.1%
DO agreement	93.3%	72.7%	70.0%
NR agreement	100.0%	100.0%	100.0%
PR agreement	100.0%	0.0%	0.0%

Table 1: Agreement figures for overall, discourse-new (DN), discourse-old (DO), non-referring (NR) and property (PR) attributes.

4 Results

The first public version of *Phrase Detectives* went live in December 2008. 1.1 million words have been converted and made ready for annotation. Over 920 players have submitted more than 380,000 annotations and validations of anaphoric relations. 46 documents have been fully annotated, meaning that at least 8 players have expressed their judgment on each markable, and each distinct anaphoric relation that these players assigned has been checked by four more players.

To put this in perspective, the GNOME corpus, produced by traditional methods, included around 3,000 annotations of anaphoric relations (Poesio, 2004a) whereas OntoNotes[5] 3.0, with 1 million words, contains around 140,000 annotations.

4.1 Agreement on annotations

A set of tools were developed to examine the decisions of the players, and address the following questions:

- How do the collective annotations produced by the game compare to annotations assigned by an expert annotator?
- What is the agreement between two experts annotating the same texts?

The answer to the first question will tell us whether the game is indeed successful at obtaining anaphoric annotations collaboratively within the game context. Anaphoric annotations are however considered much harder than other tasks such as part-of-speech tagging. Therefore we ask the second question which will give us an upper bound of what can be expected from the game in the best possible case.

We analysed five completed documents from the Wikipedia corpus containing 154 markables. We first looked at overall agreement and then broke it down into individual types of anaphoric relations. The following types of relation can be assigned by players:

- DN (discourse-new): this markable has no anaphoric link to any previous markable.
- DO (discourse-old): this markable has an anaphoric link and the player needs to link it to the most recent antecedent.
- NR (non-referring): this markable does not refer to anything e.g. pleonistic ”it”.
- PR (property attribute): this markable represents a property of a previously mentioned markable.

DN is the most common relation with 70% of all markables falling in this category. 20% of markables are DO and form a coreference chain with markables previously mentioned. Less than 1% of markables are non-referring. The remaining markables have been identified as property attributes.

Each document was also manually annotated individually by two experts. Overall, we observe 84.5% agreement between Expert 1 and the game and 83.9% agreement between Expert 2 and the game. In other words, in about 84% of all cases the relation obtained from the majority vote of non-experts was identical to the one assigned by an expert. Table 1 gives a detailed breakdown of pairwise agreement values.

The agreement between the two experts is higher than between an expert and the game. This on its own is not surprising. However, an indication of the difficulty of the annotation task is the fact that the experts only agree in 94% of all cases. This can be seen as an upper boundary of what we might get out of the game.

Furthermore, we see that the figures for DN are very similar for all three comparisons. This seems to be the easiest type of relation to be detected.

[5] `http://www.ldc.upenn.edu`

DO relations appear to be more difficult to detect. However if we relax the DO agreement condition and do not check what the antecedent is, we get agreement figures above 90% in all cases: almost 97% between the two experts and between 91% and 93% when comparing an expert with the game. A number of these cases which are assigned as DO but with different antecedents are actually coreference chains which link to the same object. Extracting coreference chains from the game is part of the future work.

Although non-referring markables are rare, they are correctly identified in every case. We additonally checked every completed markable identified as NR in the corpus and found that there was 100% precision in 54 cases.

Property (PR) relations are very hard to identify and not a single one resulted from the game.

4.2 Disagreement on annotations

Disagreements between experts and the game were examined to understand whether the game was producing a poor quality annotation or whether the markable was in fact ambiguous. These are cases where the gold standard as created by an expert is not the interpretation derived from the game.

- In 60% of all cases where the game proposed a relation different from the expert annotation, the expert marked this relation to be a possible interpretation as well. In other words, the majority of disagreements are not false annotations but alternatives such as ambiguous interpretations or references to other markables in the same coreference chain. If we counted these cases as correct, we get an agreement ratio of above 93%, close to pairwise expert agreement.

- In cases of disagreement the relation identified by the expert was typically the second or third highest ranked relation in the game.

- The cumulative score of the expert relation (as calculated by the game) in cases of disagreement was 4.5, indicating strong player support for the expert relation even though it wasn't the top answer. A relation with a score of zero would be interpreted as one that has as many players supporting it as it has players disagreeing.

4.3 Discussion

There are very promising results in the agreement between an expert and the top answer produced from the game. By ignoring property relations and the identification of coreference chains, the results are close to what is expected from an expert. The particular difficulty uncovered by this analysis is the correct identification of properties attributes.

The analysis of markables with disagreement show that some heuristics and filtering should be applied to extract the highest quality decisions from the game. In many of the cases the game recorded plausible interpretations of different relations, which is valuable information when exploring more difficult and ambiguous markables. These would also be the markables that automatic anaphora resolution systems would have difficulty solving.

The data that was used to generate the results was not filtered in any way. It would be possible to ignore annotations from users who have a low rating (judged when players annotate a gold standard text). Annotation time could also be a factor in filtering the results. On average an annotation takes 9 seconds in Annotation Mode and 11 seconds in Validation Mode. Extreme variation from this may indicate that a poor quality decision has been made.

A different approach could be to identify those users who have shown to provide high quality input. A knowledge source could be created based on input from these users and ignore everything else. Related work in this area applies ideas from citation analysis to identify users of high expertise and reputation in social networks by, e.g., adopting Kleinberg's HITS algorithm (Yeun et al., 2009) or Google's PageRank (Luo and Shinaver, 2009).

The influence of document type may have a significant impact on both the distribution of markable types as well as agreement between experts and the game. We have only analysed the Wikipedia documents, however discourse texts from Gutenberg may provide different results.

5 Conclusions

This first detailed analysis of the annotations collected from a collaborative game aiming at a large anaphorically annotated corpus has demonstrated that high-quality natural language resources can be collected from non-expert users. A game approach can therefore be considered as a possible alternative to expert annotations.

We expect that the finally released corpus will apply certain heuristics to address the cases of disagreement between experts and consensus derived from the game.

6 Future Work

This paper has focused on percentage agreement between experts and the game output but this is a very simplistic approach. Various alternative agreement coefficients have been proposed that correct for chance agreement. One such measure is Cohen's κ (Cohen, 1960) which we are using to perform a more indepth analysis of the data.

The main part of our future work remains the creation of a very large annotated corpus. To achieve this we are converting source texts to include them in the game (our aim is a 100M word corpus). We have already started converting texts in different languages to be included in the next version of the game.

Acknowledgments

ANAWIKI is funded by a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/F00575X/1. Thanks to Daniela Goecke, Nils Diewald, Maik Stührenberg and Daniel Jettka (University of Bielefeld), Mark Schellhase (University of Essex) and all the players who have contributed to the project

References

P. N. Bennett, D. M. Chickering, and A. Mityagin. 2009. Learning consensus opinion: mining data from a labeling game. In *Proceedings of the 18th International World Wide Web Conference (WWW2009)*, pages 121–130, Madrid.

L. Burnard. 2000. The British National Corpus Reference guide. Technical report, Oxford University Computing Services, Oxford.

J. Chamberlain, M. Poesio, and U. Kruschwitz. 2008. Phrase Detectives - A Web-based Collaborative Annotation Game. In *Proceedings of I-Semantics*, Graz.

J. Chamberlain, M. Poesio, and U. Kruschwitz. 2009. A new life for a dead parrot: Incentive structures in the Phrase Detectives game. In *Proceedings of the Webcentives Workshop at WWW'09*, Madrid.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL06*.

N. L. Johnson, S. Rasmussen, C. Joslyn, L. Rocha, S. Smith, and M. Kantor. 1998. Symbiotic Intelligence: Self-Organizing Knowledge on Distributed Networks Driven by Human Interaction. In *Proceedings of the Sixth International Conference on Artificial Life*. MIT Press.

D. Jurafsky and J. H. Martin. 2008. *Speech and Language Processing- 2^{nd} edition*. Prentice-Hall.

M. Krause and H. Aras. 2009. Playful tagging folksonomy generation using online games. In *Proceedings of the 18th International World Wide Web Conference (WWW2009)*, pages 1207–1208, Madrid.

M. Krötzsch, D. Vrandečič, M. Völkel, H. Haller, and R. Studer. 2007. Semantic Wikipedia. *Journal of Web Semantics*, 5:251–261.

U. Kruschwitz, J. Chamberlain, and M. Poesio. 2009. (Linguistic) Science Through Web Collaboration in the ANAWIKI Project. In *Proceedings of WebSci'09*, Athens.

X. Luo and J. Shinaver. 2009. MultiRank: Reputation Ranking for Generic Semantic Social Networks. In *Proceedings of the WWW 2009 Workshop on Web Incentives (WEBCENTIVES'09)*, Madrid.

M. Poesio. 2004a. Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*.

M. Poesio. 2004b. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*.

P. Singh. 2002. The public acquisition of commonsense knowledge. In *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, Palo Alto, CA.

M. Stührenberg, D. Goecke, N. Diewald, A. Mehler, and I. Cramer. 2007. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the ACL Linguistic Annotation Workshop*, pages 140–147.

J. Surowiecki. 2005. *The Wisdom of Crowds*. Anchor.

L. von Ahn, R. Liu, and M. Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of CHI '06*, pages 55–64.

L. von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

C. A. Yeun, M. G. Noll, N. Gibbins, C. Meinel, and N. Shadbolt. 2009. On Measuring Expertise in Collaborative Tagging Systems. In *Proceedings of WebSci'09*, Athens.