Automatic Extraction of Contextual Valence Shifters

Noémi Boubel   **Thomas François**   **Hubert Naets**

Cental, ILC (Université catholique de Louvain)

{noemi.boubel,thomas.francois,hubert.naets}@uclouvain.be

Abstract

In opinion mining, many linguistic structures, called contextual valence shifters, may modify the prior polarity of items. Some systems of sentiment analysis have tried to take these shifters into account, but few studies have focused on the identification of all these structures and their impact on polarized words.

In this paper, we describe a method that automatically identifies contextual valence shifters. It relies on a chi-square test applied to the contingency table representing the distribution of a candidate shifter in a corpus of reviews of various opinions. The system depends on two resources in French – a corpus of reviews and a lexicon of valence terms – to build a list of French contextual valence shifters. We also introduce a set of rules used to classify the extracted contextual valence shifters according to their impact on polarized words. They make use of the Pearson residuals in contingency tables to filter candidate shifters and classify them. We show that the technique reaches an F-measure of either 0.56 or 0.66, depending on how the categories of shifters are defined.

1 Introduction and State of the Art

Most opinion mining systems rely on the extraction of sentiment words to detect opinions. These words, which we will rather refer to as polarized words, convey useful information about the semantic orientation (positive or negative) of a text. However, the context in which these words appear may modify their valence in many ways. Although being of importance, this issue has been investigated only recently and is now the object of an increasing attention.

Polanyi and Zaenen (2004) first postulated the existence of *contextual valence shifters*, which are contextual phenomena altering the prior polarity of a term. Afterwards, some of these phenomena (such as negative or conditional syntactic structures) were dealt with on a case by case basis (Das and Chen, 2001; Na et al., 2004; Popescu and Etzioni, 2005; Pang et al., 2002; Wilson et al., 2005; Wilson et al., 2006; Councill et al., 2010). Studies addressing the phenomenon as a whole flourished later. They aimed at best modelling the expression of opinions (Polanyi and Zaenen, 2004; Taboada et al., 2011; Hatzivassiloglou and Wiebe, 2000; Morsy and Rafea, 2012; Musat and Trausan-Matu, 2010), before embedding those in a classification system. The main purposes of these studies are to determine a list of contextual valence shifters that impact the polarity of a term as well as to define the nature of this impact. However, these lists are often manually built from linguistic intuitions and not learned from language data. Works relying on a corpus of texts to develop resources that best reflect the actual role played by the linguistic context for opinion mining are few. Li et al. (2010) suggested a technique to automatically select polarity-shifting features in order to improve a sentiment classification system based on a machine-learning approach.

All these studies agree that contextual valence shifters can have diverse impacts on polarized words. They classify them according to the nature of this impact (Polanyi and Zaenen, 2004; Quirk et al., 1985; Kennedy and Inkpen, 2006): *inversers* invert the polarity of a polarized item, *intensifiers* intensify it and *attenuators* diminish it.

This study, based on a French corpus, focuses on the issue of contextual valence shifters and pursues two main objectives: (1) propose an automatic method that efficiently models contextual valence shifters, with the aim of improving performance of opinion mining systems (especially

those based on a term-counting method); (2) clarify the linguistic structures constituting a hindrance to current classification systems. From these two perspectives, our approach differs from the work of Li et al. (2010). Moreover, we are interested in describing the effect of all kind of modifiers (inversers, but also intensifiers and attenuators). We restricted our study to all lexico-syntactic patterns located in the immediate context of a polarized term and impacting the valence of this term. This restriction means dealing with individual words. However, it should be noted that contextual shifters may sometimes be phrases too. Our approach also relies on the assumption that contextual shifters are in direct syntactic relation with the polarized word, which has to be confirmed.

Based on the results of previous works (Boubel, 2012; Boubel and Bestgen, 2011), we propose here a system that automatically extracts modifiers (in the form of lexico-syntactic patterns) and classifies them according to their semantic impact. The general methodology is detailed in Section 2 and we report the evaluation of the method in Section 3. The paper concludes with Section 4, discussing some issues we faced, in particular the problem of the attenuating valence shifters.

2 Methodology

2.1 Key principle

In order to identify valence shifters along with their semantic impact on polarized words, we propose to exploit two different pieces of information regarding the expression of polarity in a text: (1) the overall polarity t of the text, i.e. the score assigned to it on a scale from very negative to very positive, and (2) the polarity p (positive or negative) of a polarized word which appears in the text. We noticed that the distribution of the patterns related to polarized words (i.e. potential modifiers) is influenced by the values of p and t. Intuitively, we can consider three cases:

- patterns in which p is of opposite polarity than t will mitigate or reverse the valence of their associated term;

- patterns that reinforce the polarity of a word will appear especially when p and t share the same polarity;

- finally, a larger number of expressions having an attenuating effect on p will be found

when t is around the middle of its scale (texts presenting a nuanced view).

2.2 The system

Based on this principle, we developed a system able to automatically detect and classify modifiers. It relies on two resources: (1) a corpus containing evaluative texts whose global polarity t is known and (2) a lexicon of terms whose polarity p is also known.

Our system performs a two-fold process. First, applying a parser to a corpus, we extract all syntactic dependency relationships that links a polarized term with another term (see Section 2.3). A statistical analysis is then performed to detect, among those, valence shifter candidates (see Section 2.3).

In the second step (see Section 2.4), a rule-based classifier further removes bad candidates and assigns a label to remaining modifiers that should correspond to their impact on polarized terms.

2.3 Statistical processing

In order to identify valence shifter candidates using statistical tests, the initial corpus – made up of evaluative texts whose polarity t is known – is first processed by a syntactic parser to obtain the list of all syntactic dependency relationships including a polarized term. Such relationships take the form of a pair of words (the polarized term and the candidate modifier), along with the nature of this relation (*e.g. NP(<NOM:déception>,<ADJ:total>)*). For each element of the list, three pieces of information are available: (1) the pattern itself, (2) the valence p of the term included in the structure, and (3) the score t of the text.

Then, we generalize over the relationships extracted, removing the polarized term and keeping only the valence shifter candidate and the syntactic relation linking it to its polarized term (*e.g. NP(<NOM:>,<ADJ:total>)*). This allows us to determine the frequency of each of these patterns in our corpus, in relation to two variables: the type of the pattern and the score t of the text. Based on these two variables, we build a contingency table for the patterns associated with positive terms and a second table for patterns in the context of a negative term [1].

Then, for a given pattern g, we compute a chi-

---

[1] We only keep patterns with a frequency higher than 20.

square test (Agresti, 2002) [2] where the distribution of g over the five possible values of t is compared with the distribution of all patterns except g. The chi-square value obtained is then used to decide whether the distribution of pattern g in the evaluative texts (t) is independent from the type of pattern. When the chi-square score is significant (based on a threshold α_1), we consider the pattern as a valuable valence shifter candidate.

Table 1 examplifies this analysis for the adjective *total* modifying a positive noun (*e.g.* "C'est une *réussite totale*.", *it is a total success.*). This pattern gets a chi-square of 139.67 ($p < 0.001$) and it stands out even more clearly when associated to a negative noun ($\chi^2 = 741.35$; $p < 0.001$), which confirms its interest as a good valence shifter candidate (*e.g.* "*déception totale*.", *a total disappointment.*).

2.4 Validation and classification of the candidates

At the end of our first step, we obtain a list of valence shifter candidates, selected on the basis of their chi-square score. In the second phase of our method, we apply rules primarily to identify the impact of each candidate on valence terms, but also to further filter the candidate list.

The idea is to rely on the adjusted residuals (Agresti, 2002), computed for the two contingency tables available for a candidate pattern (with negative and positive terms). Adjusted residuals corresponds to a z-score, and high values (based on a threshold α_2) means that the pattern g is either over-represented in texts with a given value of t, or is under-represented. These residuals can sometimes display specific and interesting patterns of under-representation or over-representation throughout the range of scores t possible for the texts. In previous work (Boubel, 2011), we analyzed the distributions of the adjusted residuals and we identified three typical profiles. Then, we were able to connect these profiles with their semantic role in the language, distinguishing three groups of modifiers: (1) "intensifiers", (2) "inversers", and (3) "concessive structures".

These findings were translated into a set of rules that automatically classify valence shifter candidates according to their impact on polarized terms.

Rules are based on the patterns of over-/under-representation and assign a score for each of the three classes of modifiers described above. At this stage, it is possible to apply a filtering threshold fs to remove the patterns that received a low score for all classes.

We can summary the whole set of rules as the three following trends :

1. Structures that are over-represented in situations where the valence of p is similar to that of t, regardless of the nature of the term polarity p (positive or negative), obtain a high score in the intensification category;

2. Structures that are over-represented in situations where p is the opposite of t obtain a high score in the inversion category (attenuating or an inversing role);

3. Finally, structures over-represented in reviews reporting a nuanced view (*e.g.* when $t = 3$ for texts rated on a scale from 1 to 5) obtain a high score in the concession category.

Following this method, the adjective "total" modifying a noun phrase is given a score of 8 as an "intensifier", 0 as an "inverser" and 2 as a "concessive". It is indeed under-represented with a positive noun while the text is negative and over-represented while it is positive (see Table 1). As a consequence, this pattern is classified as an intensifier.

It is worth noting that the classification underlying this approach does not match the one commonly used in the field, which draws a distinction between intensifiers, shifters, and diminishers. Our second category "inversers" includes both shifters and diminishers, since these two classes have similar statistical properties according to our method. On the contrary, the analysis of the statistical behavior of some valence shifter candidates highlights a particular semantic behavior which is not dealt with as such in the literature: it corresponds to patterns connecting several polarized terms of different polarities and having an impact on the polarity value of the whole expression. These are the patterns gathered in the third category: the "concessive structures". We observe that using statistical properties from the contingency tables to identify categories of valence shifters has limitations in terms of qualitative approach of the

---

[2]We used chi-square test as a first approach. However, it would be valuable to try other statistical tests in the future.

Score of texts (from 1 to 5) (*t*)	1	2	3	4	5	
[total-positive noun]	21 (0.74)	24 (-3.90)*	66 (-6.09)*	400 (4.65)*	536 (11.10)*	1,047
other patterns with positive noun	283,069	588,073	1,507,934	5,454,541	4,188,908	12,022,525
	283,090	588,097	1,508,000	5,454,941	4,189,444	12,023,572

Table 1: A contingency table for the adj. *total*. The adjusted residuals are significant for $\alpha_2 = 0.05$

task, but also helps to uncover interesting phenomena. We will come back to the insightful of this classification further in the paper.

3 Evaluation

The evaluation of our technique was carried out according to three steps. First, we collected the resources required by the approach, namely a corpus of evaluative texts classified according to their judgment (*t*), a valence lexicon, and a list of dependencies relationships in which modifiers have been annotated (our gold standard). They are further described in Section 3.1. Then, we carried out a quantitative evaluation of the technique, comparing its predictions to our gold standard (see Section 3.2). Finally, in Section 3.3, we conducted a qualitative analyse of the results, in order to better understand the way our technique works.

3.1 Resources

To implement our approach, the first resource needed is a corpus of texts ranked according to the opinion they express (*t*). The corpus we used was provided by the *NOMAO* company [3], which proposes a web and mobile application helping people to find, share and discover new places. It is made of 2,200,000 internet user reviews in French relative to restaurants or hotels (7,571,730 sentences). Every text has been given a score from 1 (very bad) to 5 (very good) by the author of the text.

The second resource needed is a valence lexicon, in which the polarities *p* of words are labelled. *NOMAO* also provided us with a such lexicon. It has been manually built and it includes 3,683 polarized French words relative to the domain of restaurant reviews (2,425 negative words and 1,258 positive words).

Finally, for evaluation purposes, a gold standard "corpus" was required, in which dependencies relationships containing a polarized words and a contextual valence shifter have been annotated. Since, there was no such corpus available, we randomly selected 500 sentences from the whole *NO-*

MAO corpus and discarded them from this corpus, that was therefore considered as the training corpus. The 500 sentences contained about 2,000 dependency relationships including a polarized word [4]. These relationships were manually annotated with a two-fold procedure: (1) decide whether the term associated to a polarized word is a contextual valence shifter or not, and (2) describe its impact on the polarized word, according to one of the available categories.

Regarding the categories, we decided to use a finer-grained system than the one based on statistical properties (see Section 2.4), because the category of *attenuators*, introduced in previous studies, intuitively stood out. This allowed us to discuss in Section 4 the relevance of the concession category we had statically identified. We therefore defined the four following classes: (1) intensifiers (INT) emphasize the valence of their associated term; (2) inversers (INV) inverse the valence of their associated term; (3) attenuators (ATT) mitigate the valence of their associated term; and (4) concessives (CONC) articulate terms or phrases of opposite polarities.

The list of dependency relationships were annotated by two experts in accordance with these four categories. In order to estimate their inter-rater agreement, we computed the Fleiss' kappa (Fleiss, 1971) and obtained a substantial agreement ($kappa = 0.716$) for the annotations. Finally, this corpus was equally divided into a development set – used to select the best set of parameters – and a test set, to assess the performance of the best model.

3.2 Results

Regarding the evaluation, the first issue was to define an adequate evaluation metric, since the task is a multiclass case. We opted for two different

---

[3]http://fr.nomao.com/

[4]It is worth noting that each relationship was considered in the context of the sentence it was extracted from. Therefore, a pattern repeated in the gold standard could be annotated in more than one way. Moreover, since we only dealt with the structures that our methodology can extract, modifiers not syntactically related with a polarized word were not annotated.

approaches commonly used in the literature. The first split the problem into a detection problem and a classification problem. It computes classic measures such as precision, recall, and F-measure (to which we will refer to as the F-measure 1) regarding the model's ability to detect a modifier, whatever its label. Then, the classification rate is computed through conditional accuracy (Abney, 2008). The second approach consists in computing the precision, recall and F-measure for each category independently, before averaging them to obtain a global estimation (we will refer to as the F-measure 2).

Another issue was the slight discrepancy between the set of labels from the manual annotation and the models. Manual annotation uses INT, ATT, INV, CONC, while the automatic classification uses INT, INV, CONC. For evaluation purposes, we had to project the four-label system onto the three-class one, considering that the category ATT (attenuator) was included into the category INV (inverser) (as it is already supposed in Section 2.4).

Once these two problems were sorted out, we had to perform an optimization step. Three meta-parameters can indeed be manipulated: α_1, α_2, and fs. α_1 is the criterion for the selection of candidate modifiers, since it determines the significance level of the chi-square test. α_2 is the significance threshold for the residuals; decreasing it makes it more difficult for a given structure to match a classification rule. fs is the filtering score assigned for each structure.

In order to limit the number of experiments, the following values were tested for both α_1, α_2: 0.0001, 0.0005, 0.001, 0.005, 0.01, and 0.05, while fs was kept constant ($fs \geq$ **5**). Once the best model according to α_1, α_2 was selected, values ranging from 5 to 9 were experimented for fs. The evaluation metric for all models were computed as follows: a list of modifiers included in a dependency relationship were extracted from the training corpus and used to classify the relationships from the development set. It appeared that the optimal parameters are $\alpha_1 = 0.05$ and $\alpha_2 = 0.005$, as long as we want to exploit the whole training corpus.

These optimal parameters were used to select 10,503 patterns, whose chi-square scores were significant among a total of 328,308 patterns. Then, the application of our classification rules further filtered those patterns, yielding a list of 6,612 contextual valence shifter candidates: 2,607 were labeled as INT, 2,677 were identified as INV, 1,328 were classified as CONC, and 216 were assigned to more than one categories [5]. However, among those candidates, only 1,147 structures received a score of 5 or higher. More strikingly, if we set fs to 9, then no more than 113 patterns are selected, among which are 66 INT and 47 INV, but no CONC.

Manipulating the filtering score fs reveals that the number of extracted valence shifters largely varies. We used the test corpus from Section 3.1, which contains 171 valence shifters (102 INT, 16 CONC and 23 INV or ATT), to estimate the recall, the precision, the conditional accuracy, and the two F-measures for our model trained on the training corpus (see Table 2).

The F-measure 1 (which represents the capacity of the model to rightly detect shifters) starts from 0.49 for patterns with a score of 5 or higher and reaches 0.64 when $fs \geq 9$. This corresponds to a recall of 0.86 and a precision of 0.37. It is obvious that our system considers too many patterns as valence shifters. This F-measure can however be improved if we use a stricter filtering score. It appears that the chi-square is less efficient than the classification rules to filter valence shifters.

The F-measure 2 is globally better than F-measure 1 and reaches 0.57 when filtering the patterns with intermediate scores. Interestingly, it decreases strongly for $fs \geq 9$. This can be explained by the fact that the system extracts less "concessive structure" and globally assigns a lower score to that type of structure. Only 6 CONC patterns are correctly classified for $fs \geq 5$ and the system does not detect any patterns of this type when $fs \geq 9$. As a result, the recall and precision for this category equals 0.

Finally, it is worth noting that the system obtains a very good conditional accuracy (85.9 for $fs \geq 5$ and 97.6 for $fs \geq 9$). This is a very interesting finding, since it shows that the classification rules we developed are relevant.

3.3 Qualitative analysis

To further analyze the efficiency of our extraction method, we submitted the list of the 260 shifters with a score of 8 or higher to a qualitative evalua-

---

[5]When the score used for filtering is low, a few structures can receive a same score for two classes. However, these cases disappear as soon as we filter with a score of 5.

Score (fs)	≥ 5	≥ 6	≥ 7	≥ 8	≥ 9
F-Measure 1 (recall, prec.)	0.49(.86, .34)	0.51(.84, .37)	0.55(.82, .42)	0.52(.67, .43)	0.64(.60, .69)
Conditional accuracy	85.9%	85.5%	86%	92.5%	97.6%
F-Measure 2 (recall, prec.)	0.56(.51.62)	0.55(.50.62)	0.56(.50.63)	0.49(.38.69)	0.40(.32.56)

Table 2: Evaluation measures for the model with filtering scores ranging from 5 to 9.

tion. The analysis confirms the conclusions drawn above: the system tends to consider too many patterns as shifters, but most of the actual shifters get the correct label, according to experts judgment. After cleaning manually the list, it appears that the system has correctly classified 85 patterns among 260, most of them being incorrectly recognized as valence shifters. Some limitations of our method could explain these errors.

First, it happens that the object of the judgment, also associated with polarized words, is extracted (*e.g.* NP(<ADJ:>,<NOM:accueil>)).

Second, grammatical words, such as articles, auxiliary verbs, etc. tend to be captured by the system because they are very frequent in texts. Most of these patterns are not relevant, but some others are important to extract because they can negate or reverse the valence of a polarized word (e.g. NP(<NOM:>,<_DET:aucun>)).

Also, the choice of using syntactic dependency relationships entails some limitations: the expression acting as the valence shifter is sometimes not extracted as a wole. Moreover, parsing errors frequently happen, extracting wrong patterns.

Finally, it happens that some words incorrectly recognized as valence shifters are actually polarized words missing from the valence lexicon.

To conclude this analysis, some characteristics emerge out of the correctly-classified structures. On the one hand, intensifiers (mostly adverbs and adjectives) often have a direct semantic impact on the polarized word to which they are related. On the other hand, the patterns belonging to the INV and CONC categories are more complex and heterogeneous (*e.g.* AP(<ADJ:loin de>,<ADV:>)) and often impact a phrase or a whole sentence, not directly a lexical item. As a consequence, the effect can be hard to model and it is sometimes difficult to distinguish between the patterns from these two classes, either manually or automatically.

4 Discussion and conclusion

In this paper, a new methodology for the automatic extraction and classification of valence shifters has been proposed. It reaches a very good accuracy for the classification, although it tends to extract too many structures. An interesting side of the method lies in its ability to identify relevant structures that are often not considered in other studies. In further work, it will be necessary to integrate the lexicon we obtained into a sentiment analysis system to check whether or not taking modifiers into may improve the performance.

Beyond this applicative goal, our methodology also stressed issues in the categories used to organize contextual valence shifters. The class of diminishers (or downtoners), as it is commonly referred to in the opinion mining domain, is difficult to capture in an automatic way. In our system, we defined three classes of shifters on the basis of three different statistical profiles. The INV class includes both diminishers and inversers, since their statistic profiles are very similar. The CONC class contains structures that often relates terms with different polarities. However, it is worth considering that diminishers are often used in concessive or rhetorical structures and assign them to the class CONC rather than to the class INV. The F-measure 2 for our model in this condition is interestingly better than the one reported above: 0.66 instead of 0.56 for the structures kept when $fs \geq 5$.

In view of these results, it appears that ATT can belong either to the INV class or to the CONC. Our assumption on this matter is that there is actually two types of diminishers: (1) diminishers modifying the valence of a single lexical item, that have statistical profiles closer to the INV category, and (2) diminishers used in concessive structure to attenuate the overall polarity of a phrase or a sentence, which should be included in the CONC class. This hypothesis will be tested in further work, through the analysis of the statistical profiles of manually annotated diminishers.

References

S.P. Abney. 2008. *Semisupervised Learning for Computational Linguistics*. Chapman and Hall/CRC, Ann Arbor, U.S.

A. Agresti. 2002. *Categorical Data Analysis. 2nd edition*. Wiley-Interscience, New York.

Noémi Boubel and Yves Bestgen. 2011. Une procédure pour identifier les modifieurs de la valence affective d'un mot dans des textes. In *Actes de TALN11*, volume 2, pages 137–142, Montpellier.

N. Boubel. 2011. Extraction automatique de modifieurs de valence affective dans un texte. Étude exploratoire appliquée au cas de l'adverbe. In *Travaux du Cercle belge de Linguistique*, volume 6.

N. Boubel. 2012. Construction automatique d'un lexique de modifieurs de polarité. In *Actes de TALN12*, Grenoble.

Isaac G Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 51–59. Association for Computational Linguistics.

S. Das and M. Chen. 2001. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference*, pages 37–56.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

V. Hatzivassiloglou and J. M Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305.

A. Kennedy and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

S. Li, S.Y.M. Lee, Y. Chen, C.R. Huang, and G. Zhou. 2010. Sentiment classification and polarity shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 635–643. Association for Computational Linguistics.

S. Morsy and A. Rafea. 2012. Improving document-level sentiment classification using contextual valence shifters. *Natural Language Processing and Information Systems*, pages 253–258.

C. Musat and S. Trausan-Matu. 2010. The impact of valence shifters on mining implicit economic opinions. *Artificial Intelligence: Methodology, Systems, and Applications*, pages 131–140.

J.C. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou. 2004. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. *Advances in Knowledge Organization*, 9:49–54.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86.

L. Polanyi and A. Zaenen. 2004. Contextual valence shifters. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 106–111.

A.M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Association for Computational Linguistics.

R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, and D. Crystal. 1985. *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99.