Following the Common Thread Through Word Hierarchies

Matthias J. Feiler(✉)

University of Zurich, Zürich, Switzerland
matthias.feiler@uzh.ch

Abstract. In this paper we develop a new algorithm for automatic taxonomy construction from a text corpus. In contrast to existing work, our objective is not to develop a general purpose lexicon or ontology but to identify the structure in a time–ordered sequence of documents. The idea is to identify "lead" words by which we are able to follow the common thread in the public discourse on a specific topic. Our taxonomy represents the backbone of the discourse (including names of protagonists and places) and may change over time. It is thus less rigid and universal than a lexicon and instead targets relationships that are valid in a given context. We present an example to illustrate the idea.

Keywords: Taxonomy learning · Topic tracking · On-line discourse

1 Introduction

Public attention to a topic has been shown to evolve in cycles [4]. Being able to determine the current phase within a cycle is useful both from an analytical and a practical viewpoint. Our motivation comes from finance where the maturity of a topic provides an indication of the extent to which relevant information has been *priced in*, i.e. reflected in the prices of traded assets. The concrete objective is to track the creation and evolution of themes in financial blogs. Blog conversations are by nature asynchronous and fragmented. The devices available in direct conversations for coordinating turns–of–talk [20] are not present. Participants need to track topical markers in order to follow the thread of a discussion. Our aim is to identify a temporary structure (a taxonomy) that supports the coherence of ideas and the emergence of a theme over many blogs.

One of the prerequisites for the formation of an over–arching stream of ideas is that blog participants are able to "connect the dots". Stories of universal truth relate to widely shared values or commonly understood situations. Being able to see the relation requires some degree of abstraction as the concrete format in which the story is told is unlikely to be identical over time. In a series of studies it has been shown that the human mind refers to a story in terms of a *schema* that contains abstract knowledge about a situation [9]. What are the word hierarchies that make certain schemas salient in the reader's mind? While fictional stories

P. Perner (Ed.): MLDM 2018, LNAI 10934, pp. 149–158, 2018.
https://doi.org/10.1007/978-3-319-96136-1_13

often follow pre–defined scripts we are unlikely to find a "screenplay" in online–blogs. By contrast, the common thread seems to emerge spontaneously out of nothing. Moreover, blog communication is subject to social influence [23]: a path dependency arises if an initial exchange of ideas is deemed relevant by some bloggers which subsequently form an in–group of people sharing those ideas. Later entrants may be forced to adjust their own contributions in order to conform with the existing in–group. This significantly affects the universality of the learned taxonomy.

2 Related Work

A large body of literature exists on the problem of automatic taxonomy construction which can be broadly classified in heuristic, rule–based (see e.g. [7,18]) or statistical methods, e.g. [3,6]. The main objective is to generate a universally valid semantic lexicon complementing manual constructions such as WordNet [17]. The motivation behind taxonomies is to be able to leverage on an existing knowledge base through the principle of inheritance: the information structure of root words (hypernyms) is transferred to subordinates (hyponyms). On the other hand, it is widely agreed that semantic relations are not unique [14] which has led to the development of domain specific taxonomies that are constructed from scratch [15,24].

A common design principle is to start by extracting hypernyms from raw text e.g. by using a bootstrapping technique that starts with a root concept and a doubly anchored dependency pattern [16]. The learned concepts give rise to a (densely connected) network which is subsequently filtered and simplified to induce a taxonomy. For example Chu-Liu/Edmond's algorithm may be used to find a spanning arborescence which, in turn, gives rise to a taxonomy based on the edge weights in the original graph [1,5]. In this work, the starting point is a co–occurrence matrix of terms computed over a domain–specific corpus. The motivation behind this choice comes conversation studies. The idea is that mutual understanding in human conversations is established "on–the–fly" through the creation of text worlds [8], i.e. shared mental representations of the situation at hand. We define text worlds as context–specific taxonomies; in fact, *local word hierarchy* would be a more suitable term for this temporary construction. The idea is to be able to follow a common thread in (public) discourse by identifying the hierarchy of keywords used in the conversation. Incidentally, higher–ranked words will correspond to more common (or abstract) ideas which brings this definition of a text world close to the notion of a taxonomy.

3 Taxonomy Generation

When people communicate, they rely on conventions in order to understand and produce meaning. Meaning is constructed in the mind of the listener using language as an input from which conceptual representations are formed. These linguistic inputs typically under–specify the concepts intended by the speaker and rely on the listener's ability to contribute the context needed to make a correct inference. In rational interaction models the speaker and listener apply (and expect) a common logic, or *cooperative principle* [11] to organize their speech acts.

The principle has been spelled out into four conversational maxims, the maxim of quality (truthfulness), quantity (informativeness), relevance and manner (conciseness). Mutual agreement on the maxims allow the speaker and listener to enrich an utterance by so–called *implicatures* which suggest an extension or modification of meaning beyond the literal interpretation, such as in *S1:* "Will he come?" *S2:* "His car broke down." which is decodable by *S1* into "He won't." by assuming that *S2* did not choose the answer if it was irrelevant. Also, *S2* supposes that *S1* has the background information that if cars brake down, people frequently do not manage to keep appointments. This is referred to as the *common ground* [2,22]. The interactive alignment model [19] emphasizes the importance of tacit coordination and *implicit* common ground. According to the model, grounding occurs automatically and the speakers' particular choices (i.e. which information to foreground) lead to an alignment of their (mental) representations.

Following a long tradition [21], conversation analysts study the way an interaction order [10] is established in practice, in particular how people take turns at talk, how they deal with overlaps and interruptions and how the sequence of utterances (and more general [speech] actions) is organized. Conversation analysis argues that the "...meaning of an action is heavily shaped by the sequence of actions from which it emerges, and that the social context is dynamically created [...] through the sequential organization of interaction", see [13], p.223. Any statement has to signal understanding of the preceding statements and prepare the floor for the next in order to establish coherence. This means that "each sentence [...] must contain some direct or indirect indication as to how it fits into the stream of talk", see [12], p.119. Two minds have to collaborate in order to "make progress" on the subject of their discussion. In the interactive alignment model this process occurs with a minimal amount of modeling what others know. According to the model, grounding occurs automatically through the speakers' particular choices i.e. which information to foreground. In this paper, we aim at identifying the words that have been foregrounded in a corpus of financial blogs.

3.1 An Algorithm for Taxonomy Extraction

Step 0 of our construction is to reduce and slice the corpus using a simple keyword filter and suitable time–intervals (e.g. a monthly grid). This generates a time–ordered sequence of sub–corpora containing documents related to a given area of interest. We represent every sub–corpus as a *bag of words* using a term–document matrix d. d is a $n \times m$ boolean matrix indicating the presence of a given term in a document where n is the number of terms (only *important* terms are included according to a global, i.e. topic–unspecific, tf–idf measure) and m is the number of documents in the sub–corpus. We aggregate over documents by setting $C = d\,d^T$ which is the co–occurrence matrix of terms in the sub–corpus.

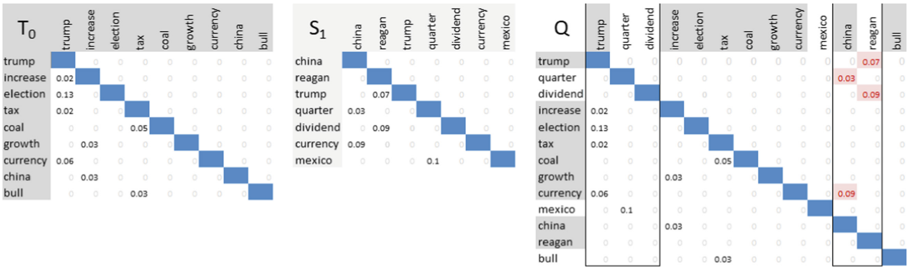
C forms the basis for the construction of our taxonomy T which proceeds in three steps:

1. Normalization of the rows of C enables us to interpret the entries of the co–occurrence matrix as intensities of a flow from every row term to the set of column terms in C. The resulting row–stochastic matrix A has a natural interpretation as the adjacency matrix of a directed graph representing the network flow originating at the term–nodes. Every edge in the resulting di–graph may be thought of as a reference that one term makes to another.
2. The nodes of the directed graph are rank–ordered by their in-degree (column sum) and the matrix A is re–arranged accordingly. We obtain: $P^T AP$ where P is the permutation matrix corresponding to the sort. Any high ranking node will be a parent node to the ones that reference it which means that the direction of any edge in the final structure is towards that node while edges out of the node are omitted. This means that the upper triangular matrix is set to zero and we obtain the intermediary result
$$T' = \mathrm{ltri}(P^T AP) \tag{1}$$
3. A *unique* parent is determined for every node by identifying the location of the maximum weight in every row of T'. We denote by $[\cdot]_{\max}$ the operator that sets all row entries except the maximum to zero and obtain
$$T = [T']_{\max} \tag{2}$$
corresponding to an in–tree or (anti–)arborescence representing a word hierarchy derived from the co–occurrence matrix. Notice that the maximum may not be unique as C is an integer matrix possibly containing duplicate entries which remain even after normalization. This is amended by choosing one of the solution candidates at random.

In summary, the number of references a term receives from others induces an order–relation which forms the backbone of our taxonomy T. In this paper, we are interested in studying how T evolves over a time–ordered sequence of sub–corpora. We introduce the index t referring to a point in the time–grid.

3.2 Taxonomy Evolution

Let T_t be a given taxonomy at instant of time t and let S_{t+1} be a new taxonomy created from the “next” sub–corpus i.e. from documents collected over the time interval $(t,\, t+1]$. Notice that S_{t+1} is an independently created taxonomy having no overlap with the previous step, i.e. the documents of T_t. The question is whether S_{t+1} may be attached to T_t in a natural way thereby extending the ideas of T_t to form a new, combined taxonomy T_{t+1}. Evolution in this context means that a given tree grows by forming branches which do not contradict the existing tree structure. While a simple keyword filter leads to an appropriate sub–corpus (for given a topic), the tree T_t specifies what is *commonly understood*



T0	trump	increase	election	tax	coal	growth	currency	china	bull
trump	■	0	0	0	0	0	0	0	0
increase	0.02	■	0	0	0	0	0	0	0
election	0.13	0	■	0	0	0	0	0	0
tax	0.02	0	0	■	0	0	0	0	0
coal	0	0	0	0.05	■	0	0	0	0
growth	0	0.03	0	0	0	■	0	0	0
currency	0.06	0	0	0	0	0	■	0	0
china	0	0.03	0	0	0	0	0	■	0
bull	0	0	0	0.03	0	0	0	0	■

S1	china	reagan	trump	quarter	dividend	currency	mexico
china	■	0	0	0	0	0	0
reagan	0	■	0	0	0	0	0
trump	0	0.07	■	0	0	0	0
quarter	0.03	0	0	■	0	0	0
dividend	0	0.09	0	0	■	0	0
currency	0.09	0	0	0	0	■	0
mexico	0	0	0	0.1	0	0	■

Q	trump	quarter	dividend	increase	election	tax	coal	growth	currency	mexico	china	reagan	bull
trump	■	0	0	0	0	0	0	0	0	0	0	0.07	0
quarter	0	■	0	0	0	0	0	0	0	0	0.03	0	0
dividend	0	0	■	0	0	0	0	0	0	0	0	0.09	0
increase	0.02	0	0	■	0	0	0	0	0	0	0	0	0
election	0.13	0	0	0	■	0	0	0	0	0	0	0	0
tax	0.02	0	0	0	0	■	0	0	0	0	0	0	0
coal	0	0	0	0	0	0.05	■	0	0	0	0	0	0
growth	0	0	0	0.03	0	0	0	■	0	0	0	0	0
currency	0.06	0	0	0	0	0	0	0	■	0	0.09	0	0
mexico	0	0.1	0	0	0	0	0	0	0	■	0	0	0
china	0	0	0	0.03	0	0	0	0	0	0	■	0	0
reagan	0	0	0	0	0	0	0	0	0	0	0	■	0
bull	0	0	0	0	0	0.03	0	0	0	0	0	0	■

Fig. 1. Attaching S_1 to T_0 through "zipping".

by the topic. We ask how documents up to instant t prepare the ground for subsequent statements which either validate the word hierarchy in T_t or propose a new one. In the former case the attachment of S_{t+1} succeeds, otherwise a new tree is started. We define the attachment operator $h : T \times S \rightarrow T$ and obtain the evolution equation:

$$T_{t+1} = h(T_t, S_{t+1}, \theta) \tag{3}$$

Together with an initial condition T_0 this equation defines a path $\{T_t\}_{t\geq 0}$. The parameter θ refers to the minimum similarity among S_{t+1} and T_t such that S_{t+1} is attached, otherwise $T_{t+1} = T_t$. The operator $h(\cdot, \cdot)$ corresponds to the following construction:

Let V be the set of columns of S_{t+1} that also appear in T_t. We re–order V according to their ranking in T_t. This will create entries in the upper triangle of the combined matrix Q. We let ξ_{t+1} be the sum of these entries normalized by the sum over all elements in S_{t+1}. Parents (i.e. higher–ranking columns) in S_{t+1} that also appear in T_t may thus become children in Q, see Fig. 1 where $t = 0$. By contrast, all columns W that do not appear in T_t retain their ranking relative to the next higher–ranking column in V. In other words, parents in S_{t+1} take their children with them as long as this does not create a contradiction with existing parents in T_t. Columns in W are inserted together with corresponding rows to form an extended, quadratic matrix Q. If $\xi_{t+1} \leq \theta$, the similarity of S_1 and T_0 is sufficient for integration and the equivalent of step three (see Sect. 3.1) is repeated on ltri(Q) in order to determine unique parents for every term (except the highest–ranking one). More precisely, the $[\cdot]_{\max}$ operator is applied only on new terms (those in W), while existing child–parent relations (those in T_t) are retained. This means that if term A is parent to term B in T_t it will continue to be parent in the new taxonomy T_{t+1}. If $\xi_{t+1} > \theta$, the ranking of columns in S_{t+1} is deemed too different than the one in T_t which means that S_{t+1} cannot be attached. The overall procedure corresponds to the "zipping" of two trees.

This is illustrated in Fig. 1: The reordering of the column "china" (top rank in S_1) creates entries in upper triangle of Q, such as the reference coming from "currency" (value: 0.09). "Currency" is subordinate to "china" in S_1 but no longer in Q as the existing tree T_0 expects a higher rank of "currency" than "china". By contrast, column "reagan" also has a lower rank than "china" in S_1 but it does not appear in T_0 which means that it can be moved to the combined structure Q together with “china”. The references for “reagan” come from “trump” and “dividend” which are both higher–ranking in Q and will therefore be set to zero. Notice that the determination of the rank occurs *before* the actual tree construction, i.e. before setting the upper triangular matrix to zero and before determining unique parents. This means that high–ranked terms (such as “reagan”) may end up with no references (if these come from even higher–ranking terms). It follows that the rank in the adjacency matrix T encodes some extra information about the taxonomy which is not reflected in the graph of T.

Comment: To some extent, the zipping operation mimics the coordination procedures in natural conversations: an existing taxonomy prepares the ground for future hierarchies to be attached (as branches). If this occurs, the top node of the sub–hierarchy is attached to a point in the taxonomy thus making a clear reference to its “origin”. In dialogue, speaker and listener adapt to each other in the sense that messages are designed to the listener (gradually incorporating the listener’s mental representations of the matter discussed) while listeners provide clear references to (or even repeat) what they heard. Over time, a chain of statement-response type of pairs (so–called adjacency pairs[1]) result which form the basis of the common thread in the dialogue. Our algorithm design draws on this basic mechanism to construct a taxonomy that evolves over time. It is clear that the analogy fails at the point where we do not consider individual conversation partners but merely aggregate text documents published over a given time period. However, if we allow ourselves to view the documents as statements of an abstract aggregate speaker or listener, the number of successful attachments indeed reflects the degree of *mutual understanding* that develops among the contributors to the text corpus.

4 Example

In Fig. 2 we illustrate the above ideas on a sub–corpus around the keyword “protectionism”. The four graphs display snapshots of the evolution of the topic taxonomy. The underlying raw data was collected in monthly batches over a time–span from May 2015 to Mar 2018. The first graph is the result of a pure taxonomy extraction from the co-occurrence matrix C obtained in May 2015. In the subsequent steps new sub–graphs are attached using Eq. (3), thereby incrementally growing the initial tree. In this illustrative example, the threshold is set to $\theta > 1$ effectively posing no constraint to the attachment process. The number of terms in C is $n = 100$ and the number of documents behind

[1] Adjacency pairs constitute the central organizing format in natural conversations. They consist of two turns by two different speakers which are relatively ordered. The so–called “first pair part” initiates the exchange whereas the “second pair part” *responds* by providing a relevant follow–up statement. In this paper, we assume that the responses are always “pair–type related”; by starting with a filtered sub–corpus we exclude improper pairings whose dialogue–equivalent would roughly read: “Would you like some tea?”–“Hi!” [21].

C varies $m = 100 \ldots 500$ depending on the intensity of the discussion around "protectionism" (i.e. the number of documents retrieved in a given month).

Notice that the study of the intensity evolution is outside the scope of this paper which is focused on the *qualitative* evolution of the topic. In fact, the entries of C (absolute occurrence counts) are normalized as part of the taxonomy extraction algorithm. In order to keep the presentation uncluttered only significant nodes are displayed in the graphs. We use an additional threshold $\theta_0 = 1.5$ for the column sum (corresponding to the aggregate endorsement received by the node) below which we do not display a node[2]. In the chart at the bottom of Fig. 2 we report the monthly values of the dissimilarity measure ξ defined above.

In May 2015, we find that our algorithm puts "trade" as a root together with qualifiers "global" and "china" which seems very close to a textbook (i.e. lexical) definition of protectionism. Around Sep 2016, near the pinnacle of the US electoral campaign, the discussion on trade has evolved to a more nuanced level containing specific issues such as "steel" and a number of macro–aspects such as "inflation rate". At the same time, a new subtree has emerged containing the "clinton/trump – scenario". Notice that the subtree has no visible connection with the protectionism discussion but has already been assigned a position within the hierarchy (see the comment at the end of Sect. 3.2). After the election, from Nov 2016 onwards, we notice an accentuated increase in the dissimilarity ξ. This marks a change in the perspective on "protectionism" which is reflected in a re–shuffle of the word–order developed thus far. In other words, new subtrees attached to the Sep 2016 tree generate more and more entries in the upper triangular matrix of the combined taxonomies. Referring to the above description of "zipping" we know that the attachment points are elements of the column set V which intersects the existing tree. The question is if these entries in the word hierarchy entail a sufficient number of sub–ordinates (i.e. elements of W having no overlap with existing structures) or even followers (i.e. sub–ordinates that also connect to terms in V). In such a case, ξ will decrease as no further contradictions are produced.

It is interesting to consider what kind of input would lead to a continuous high level of ξ: this would correspond to a sustained re–shuffling of the word order which would mean that the position of any new word introduced to the hierarchy would be revised in subsequent months. This is characteristic of a change in viewpoints or interpretations on a topic as can be seen in the period after Sep 2016. The Nov 2016 taxonomy shows that two subtrees may initially grow independently with "trump" becoming the root of the "election" tree. In Mar 2017 this tree finally connects to the "trade–china–rate" tree bringing a number of new elements into the discussion such as "mexico", "currency" and "dollar". It should be noted that the "trump" compound is sub–ordinate to the earlier discussion around the macro effects of "protectionism".

Notice also, that the structure of the final taxonomy depends on the initial condition: if the attachment process had been started at a later stage, say in Nov 2016, "trump" would have been the root. An important feature of the

[2] This level of θ_0 is thus 1.5 times the row sum in the normalized matrix C.

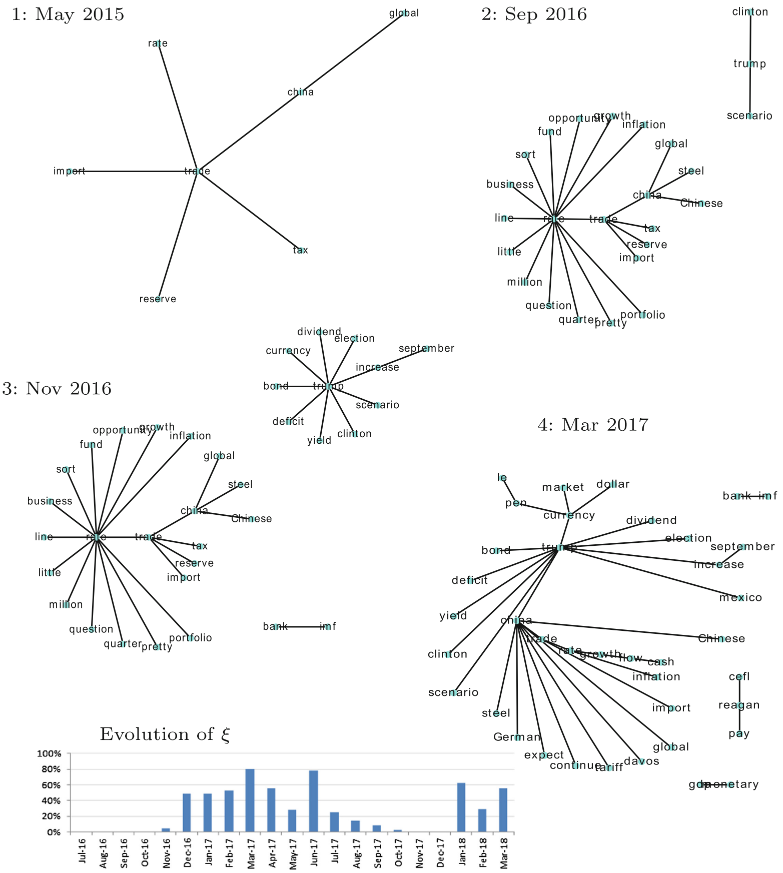


Fig. 2. Example: taxonomies of the theme "protectionism" (generated through attachment of monthly sub–corpora according to Eq. 3) and evolution of the dissimilarity measure ξ.

proposed technique is indeed that it indicates the origin of a discussion. In fact, our construction is path–dependent, as is the formation of common ground in natural conversations. After Mar 2017 we see that ξ declines indicating a steady state in the taxonomy. This temporary definition of the implications and ramifications of protectionism is again challenged in Jun 17 and Jan 18 as reflected by a resurgence of ξ.

5 Conclusion

The paper presented a new algorithm for the automatic construction of a temporary taxonomy used in on–line conversations to establish a (context–dependent) common ground. The taxonomy evolves as new sub–topics enter the conversation. A natural question about the algorithm is how it may be benchmarked.

Two taxonomies may be compared in terms of their “normative capacity”, i.e. their ability to establish a word hierarchy which attracts followers (in terms of trees attached). Given a base tree, the question is whether subsequent trees may be attached without significant changes in the word order. If ξ in the above construction is large (and remains so), the trees contain the same set of keywords but in a different order. It is then possible to search for another base which leads to a decreasing ξ as new trees are attached. This is equivalent to a gradual specification of the defining word hierarchy associated with a topic. If ξ indeed decreases, more and more following trees attach to an existing base using the same word order and adding new words which do not contradict the existing structure. Notice the *self–referential* nature of this definition: a taxonomy is “true” if it is used by many subsequent documents. This is in contrast to benchmarking against an exogenous ground truth as given by a lexicon or an established ontology. In on–line discourse, “ground truth” is a fluid concept and reflects what most people think. The fact that a taxonomy is validated *from within* – through mutual understanding among contributors – marks a departure from standard problems in taxonomy construction.

References

1. Chu, Y.J.: On the shortest arborescence of a directed graph. Sci. Sin. **14**, 1396–1400 (1965)
2. Clark, H.H., Marshall, C.R.: Definite reference and mutual knowledge. Psycholinguistics: Crit. Concepts Psychol. **414** (2002)
3. Cohen, T., Widdows, D.: Empirical distributional semantics: methods and biomedical applications. J. Biomed. Inform. **42**(2), 390–405 (2009)
4. Downs, A.: Up and down with ecology-the issue-attention cycle. Public Interest **28**, 38–50 (1972)
5. Edmonds, J.: Optimum branchings. J. Res. Natl. Bureau Stan. B **71**(4), 233–240 (1967)
6. Fountain, T., Lapata, M.: Taxonomy induction using hierarchical random graphs. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 466–476. Association for Computational Linguistics (2012)
7. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI, vol. 7, pp. 1606–1611 (2007)
8. Gavins, J.: Text World Theory. Edinburgh University Press, Edinburgh (2007)
9. Gick, M.L., Holyoak, K.J.: Schema induction and analogical transfer. Cogn. Psychol. **15**(1), 1–38 (1983)
10. Goffman, E.: Forms of Talk. University of Pennsylvania Press, Philadelphia (1981)
11. Grice, H.P.: Logic and conversation, pp. 41–58 (1975)
12. Gumperz, J.J.: Mutual inferencing in conversation. In: Mutualities in Dialogue, pp. 101–123 (1995)
13. Heritage, J.: Conversation analysis and institutional talk. In: Handbook of Language and Social Interaction, pp. 103–147 (2005)
14. Hovy, E.: Comparing sets of semantic relations in ontologies. In: Green, R., Bean, C.A., Myaeng, S.H. (eds.) The Semantics of Relationships, vol. 3, pp. 91–110. Springer, Heidelberg (2002). https://doi.org/10.1007/978-94-017-0073-3_6

15. Kozareva, Z., Hovy, E.: A semi-supervised method to learn and construct taxonomies using the web. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1110–1118. Association for Computational Linguistics (2010)
16. Kozareva, Z., Riloff, E., Hovy, E.H.: Semantic class learning from the web with hyponym pattern linkage graphs. In: ACL, vol. 8, pp. 1048–1056 (2008)
17. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: an on-line lexical database. Int. J. Lexicography **3**(4), 235–244 (1990)
18. Pantel, P., Pennacchiotti, M.: Espresso: leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of the 21st International Conference on Computational Linguistics, pp. 113–120. Association for Computational Linguistics (2006)
19. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behav. Brain Sci. **27**(02), 169–190 (2004)
20. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language 696–735 (1974)
21. Schegloff, E.A.: Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis, vol. 1. Cambridge University Press, Cambridge (2007)
22. Stalnaker, R.: Common ground. Linguist. Philos. **25**(5–6), 701–721 (2002)
23. Turner, J.C.: Social Influence. Thomson Brooks/Cole Publishing Co, Pacific Grove (1991)
24. Velardi, P., Faralli, S., Navigli, R.: Ontolearn reloaded: a graph-based algorithm for taxonomy induction. Comput. Linguist. **39**(3), 665–707 (2013)