Expanding Training Datasets by Generalization of Linguistic Structures

Boris Galitsky
Oracle Corp., Redwood Shores CA USA
and
Dmitry Ilvovsky
Higher School of Economics, Moscow, Russia

Abstract

We convert existing training datasets into the ones closed under linguistic generalization operations to expand infrequent cases. We transfer the definition of least general generalization from logical formulas to linguistic structures, from words to phrases, sentences, speech acts and discourse trees. The main advantage of the resultant frameworks is explainability and learnability from a small set of samples. Learning via generalization of linguistic structures turned out to be well suited for industrial linguistic applications with limited training datasets.

1 Introduction

A lack of data, especially covering tail phenomena, is a major bottleneck for language learning system. As statistical and deep learning language systems provide higher overall accuracy in most cases, it is never obvious how to circumscribe these successful cases and how to extend the training datasets to cover tail, unsuccessful cases (Ettinger et al., 2017., Kovalerchuk & Kovalerchuk 2017). To address this problem we expand a training dataset into a form that would force the learning framework to acquire generalizations from it.

Our learning framework includes the generalizer module that applies linguistic generalization procedure to multiply cases, which are under-represented in the original training set. This generalization procedure is deterministic and enables explainability and interpretability features of the training set. Then the expanded training set is fed to a statistical or deep learning module which is expected to generalize well beyond the original training data. Hence we intend to achieve the best of both worlds: high recognition accuracy of deep learning and generalization completeness of the training set delivered by the rule-based, interpretable generalization procedure.

2 Generalization of Texts

To measure of similarity of abstract entities expressed by logic formulas, a least-general generalization was proposed for a number of machine learning approaches, including explanation based learning and inductive logic programming. Least general generalization was originally introduced by (Plotkin 1970). It is the opposite of most general unification (Robinson 1965) therefore it is also called anti-unification. For two words of the same POS, their generalization is the same word with POS. If lemmas are different but POS is the same, POS stays in the result. If lemmas are the same but POS is different, lemma stays in the result.

Let us represent a meaning of two natural language expressions by logic formulas and then construct unification and anti-unification of these formulas. Some words (entities) are mapped into predicates, some are mapped into their arguments, and some other words do not explicitly occur in logic form representation but indicate the above instantiation of predicates with arguments. How to generalize the expressions?

- camera with digital zoom
- camera with zoom for beginners

To express the meanings we use logic predicates *camera(name_of_feature, type_of_users)* (in real life, we would have much higher number of arguments), and *zoom(type_of_zoom).* The above NL expressions will be represented as:

camera(zoom(digital), AnyUser)

camera(zoom(AnyZoom), beginner),

where variables (non-instantiated values, not specified in NL expressions) are capitalized. Given the above pair of formulas, unification computes their most general specialization *camera(zoom(digital), beginner),* and anti-unification computes their most specific generalization, *camera(zoom(AnyZoom), AnyUser).*

At syntactic level, we have generalization ('^') of two noun phrases as: {NN-camera, PRP-with, [digital], NN-zoom [for beginners]}.

We eliminate expressions in square brackets since they occur in one expression and do not occur in another. As a result, we obtain{NN-camera, PRP-with, NN-zoom]}, which is a syntactic analog as the semantic generalization above.

The purpose of an abstract generalization is to find commonality between portions of text at various semantic levels. Generalization operation occurs on the levels of Text / Paragraph / Sentence / Individual word.

At each level except the lowest one, individual words, the result of generalization of two expressions is a set of expressions. In such set, for each pair of expressions so that one is less general than other, the latter is eliminated. Generalization of two sets of expressions is a set of sets which are the results of pairwise generalization of these expressions.

Only a single generalization exists for a pair of words: if words are the same in the same form, the result is a node with this word in this form. To involve *word2vec* models (Mikolov et al., 2015), compute generalization of two different words, we use the following rule. If *subject1=subject2,* then *subject1^subject2 = <subject1, POS(subject1), 1>*. Otherwise, if they have the same part-of-speech, *subject1^subject2 =<*,POS(subject1), word2vecDistance(subject1^subject2)>*. If part-of-speech is different, generalization is an empty tuple. It cannot be further generalized.

For a pair of phrases, generalization includes all maximum ordered sets of generalization nodes for words in phrases so that the order of words is retained. In the following example

To buy digital camera today, on Monday

Digital camera was a good buy today, first Monday of the month

generalization is {*<JJ-digital, NN-camera> ,<NN- today, ADV,Monday>*} , where the generalization for noun phrases is followed by the generalization by adverbial phrase. Verb *buy* is excluded from both generalizations because it occurs in a different order in the above phrases. *Buy - digital - camera* is not a generalization phrase because *buy* occurs in different sequence with the other generalization nodes.

At the discourse level, rhetorical relations with elementary discourse units can be generalized as well. Only rhetorical relations of the same class (*presentation* relation, such as *antithesis, subject matter* relation, such as *condition, and multinuclear* relation, such as *list*) can be generalized. We use *N* for a nucleus or situations presented by this nucleus, and *S* for satellite or situations presented by this satellite. *Situations* are propositions, completed actions or actions in progress, and communicative actions and states (including *beliefs, desires, approve, explain, reconcile* and others). Hence we have the following expression for Rhetoric Structure Theory (RST, Marcu 2000) based generalization for two texts $text_1$ and $text_2$:

$text_1 \wedge text_2 = \cup_{i,j} (rstRelation_{1i,} (\ldots,\ldots) \wedge rstRelation_{2j} (\ldots,\ldots))$,

where $I \in$ (RST relations in $text_1$), $j \in$ (*RST relations in* $text_2$). Further, for a pair of RST relations their generalization looks as follows: $rstRelation_1(N_1, S_1) \wedge rstRelation_2 (N_2, S_2) = (rstRelation_1 \wedge rstRelation_2)(N_1 \wedge N_2, S_1 \wedge S_2)$.

The texts in N_1, S_1 are subject to generalization as phrases. The rules for $rst_1 \wedge rst_2$ are as follows. If $relation\_type(rst_1) \; ! = relation\_type(rst_2)$ then similarity is empty. Otherwise, we generalize the signatures of rhetoric relations as sentences: $sentence(N_1, S_1) \wedge sentence (N_2, S_2)$ (Iruskieta et al 2015).

To optimize the calculation of generalization score, we rely on a computational study which determined the POS weights to deliver the most accurate similarity measure between sentences possible

(Galitsky et al 2012). The problem was formulated as finding optimal weights for nouns, adjectives, verbs and their forms (such as gerund and past tense) such that the resultant search relevance is maximum. Search relevance was measured as a deviation in the order of search results from the best one for a given query (delivered by Google); current search order was determined based on the score of generalization for the given set of POS weights (having other generalization parameters fixed). As a result of this optimization performed in (Galitsky et al 2012), we obtained $W_{NN} = 1.0$, $W_{JJ} = 0.32$, $W_{RB} = 0.71$, $W_{CD} = 0.64$, $W_{VB} = 0.83$, $W_{PRP} = 0.35$ excluding common frequent verbs like *get/ take/set/put* for which $W_{VBcommon} = 0.57$. We also set that $W_{<POS,*>} = 0.2$ (different words but the same POS), and $W_{<*,word>} = 0.3$ (the same word but occurs as different POSs in two sentences).

Generalization score (or similarity between sentences $sent_1$, $sent_2$) then can be expressed as sum through phrases of the weighted sum through words $word_{sent1}$ and $word_{sent2}$

$score(sent_1, sent_2) = \sum_{\{NP, VP, \ldots\}} \sum W_{POS}\, word\_gen(word_{sent1}\, word_{sent2})$. The best generalization can then be defined as the one with the highest score. This way we define a generalization for phrases, sentences and paragraphs, as well as verb signatures from VerbNet, speech acts, communicative actions, as well rhetorical relations.

Result of generalization can be further generalized with other parse trees or generalization. For a set of sentences, totality of generalizations forms a lattice: order on generalizations is set by the subsumption relation and generalization score (Khardon and Arias. 2006).

3 Question Answering Relevance via Generalization

To rank answers according to the relevance to a complex paragraph-sized question, paragraph-level generalization comes into play. It needs to involve a set of parse trees for each sentence of the paragraph and the paragraph-level discourse information. We refer to the sequence of parse trees plus a number of arcs for inter-sentence relations of the discourse type between the nodes for words as a *parse thicket* (Galitsky 2017). It is a graph that includes parse trees for each sentence, as well as additional arcs for inter-sentence discourse relationships. We intend to demonstrate the richness of representation by parse thickets and the robustness of syntactic generalization operation on them for search relevance.

Our example is a web search query and its answers selected from the first page of a Google search. Although both answers *A1* and *A2* share very similar keywords, we show that using discourse information helps to differentiate them, relying on parse thicket representation and syntactic generalization operation with query *Q*.

Q: I am buying a foreclosed house. A bank offered me to waive inspection; however I am afraid I will not identify some problems in this property unless I call a specialist.
A1: *My wife and I are buying a foreclosure from a bank. In return for accepting a lower offer, they want me to waive the inspection. I prefer to let the bank know that I would not waive the inspection … Instead I would agree that I would pay the costs involved for a proper and thorough home inspection…*
A2: *I am a foreclosure specialist in a bank which is subject to an inspection. FTC offered us to waive inspection if we can identify our potential problems with customers we lent money to buy their properties.*

The reader can see that *A2* is totally irrelevant while *A1* is relevant.

We selected the first Google search result for the correct answer and composed a totally irrelevant answer with the same keywords to demonstrate the role of discourse-level generalization.

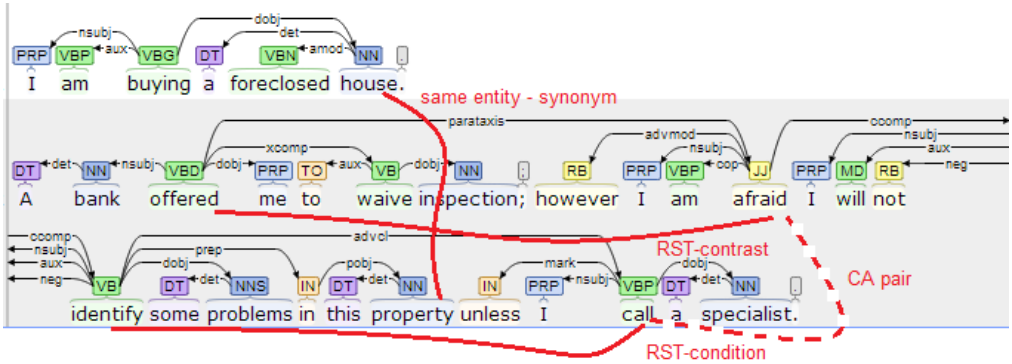

Fig. 1a parse thicket for question *Q*

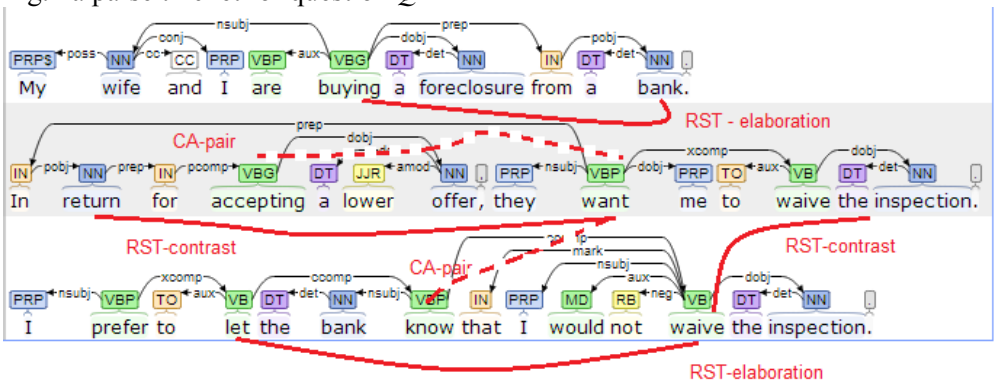

Fig. 1b Parse thicket for the valid answer *A1*

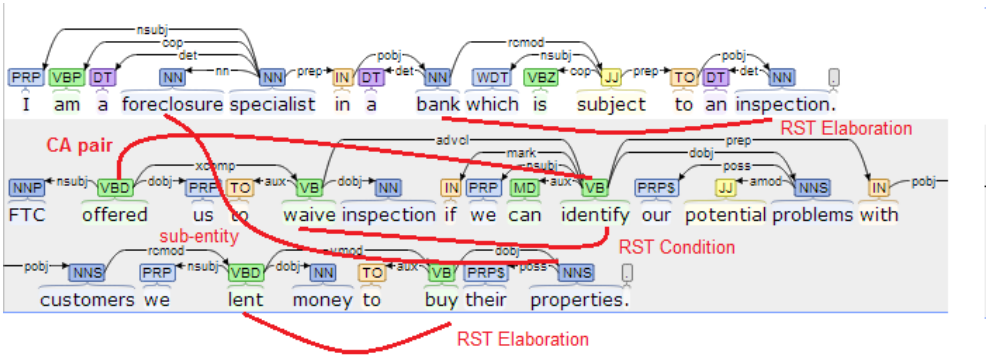

Fig. 1c: Parse thicket for the invalid answer *A2*

The list of common keywords gives us a hint that both documents are about a relationship between the same entities, a *house, a buyer* and *a bank* in connection to a *foreclosure* and an *inspection*. However one can see that the relations between these entities in *A1* and *A2* are totally different. It is also obvious that something beyond the keyword statistics and n-gram analysis needs to be done to figure out the correspondence of the structure of these relations between *A1* and *Q*, and *A2* and *Q*.

Buy, foreclosure , house, bank, wave, inspection..

One can see that the key for the right answer here is rhetorical (RST) relation of *contrast*: *bank wants the inspection waved* but *the buyer does not*. Parse thicket generalization gives the detailed similarity picture that looks more complete than both the bag-of-words approach and pair-wise sentence generalization would. The similarity between *Q* and *A1* is expressed as a parse thicket expressed here as a list of phrases

[[NP [DT-a NN-bank], NP [NNS-problems], NP [NN*-property], NP [PRP-i]], VP [VB-* TO-to NN-inspection], VP [NN-bank VB-offered PRP-* TO-to VB-waive NN-inspection], VP [VB-* VB-identify NNS-problems IN-* NN*-property], VP [VB-* {phrStr=[], roles=[A, *, *], phrDescr=[]} DT-a NN-*]]]

And similarity with the invalid answer *A2* is expressed as a parse thicket formed as a list of phrases

[[NP [DT-a NN-bank], NP [PRP-i]], [VP [VB-* VB-buying DT-a], VP [VB-* PRP-me TO-to VB-waive NN-inspection], VP [VB-* {phrStr=[], roles=[], phrDescr=[]} PRP-i MD-* RB-not VB-* DT-* NN*-*],

The important phrases of the *Q ^ A1* similarity are VP [NN-bank VB-offered PRP-* TO-to VB-waive NN-inspection], VP [VB-* VB-identify NNS-problems IN-* NN*-property],
which can be interpreted as a key topic of both *Q* and A1*: bank* and not another entity should *offer to waive inspection*. This is what differentiates *A1* from *A2* (where these phrases are absent). Although *bank* and *problems* do not occur in the same sentences in Q and A1, they were linked by anaphora and RST relations. Without any kind of discourse analysis, it would be hard to verify whether the phrases containing *bank* and *problems* are related to each other. Notice that in *A2*, problems are associated with *customers*, not *banks*, and different rhetoric relations from those common between *Q* and *A1* help us figure that out. Notice the semantic role attributes for verbs such as VB-* {phrStr=[], roles=[A, *, *], phrDescr=[]} in generalization result.

Parse thickets for *Q, A1* and *A2* are shown in Fig. 1a, 1b and 1c respectively. Notice the similarity in discourse structure of *Q, A1* and not in *A2:* the RST-*contrast* arc. Also, there is a link for a pair of communicative actions for *Q, A1* (it is absent in *A2*): *afraid-call* and *accept-want*.

4 Experiments and Conclusions

In this section we briefly enumerate a number of tasks and the results for original and extended dataset. We do not provide details of the datasets and evaluation problems and settings but only show the contribution of dataset expansion. It will give a clue on how dataset expansion with the focus of generalization helps in solving problems requiring rich semantic representation.

One can observe a 1-4% improvement in F1 for the **typical** cases (shown in bold) and 4-7% improvement for the tail cases when the dataset is expanded by the paragraph-level generalization. For some domains transition from sentence to paragraph-level generalization is beneficial.

Our conclusion is that generalization operation on the training set multiplies tail cases, makes it more balanced, and eliminates noisy samples which cannot be generalized, and the same learning algorithm delivers higher accuracy.

Problem	Original dataset	Expansion with sentence-level generalization	Expansion with paragraph-level generalization
Searching complex, multi- sentence queries	**79.1** /67.2	**83.6** / 69.3	**86.4** / 74.2
Dialogue management	**67.4**/ 60.2	**69.0** / 64.1	**72.7** / 65.8
Document style recognition	**88.3**/ 80.4	**89.3**/ 83.9	**89.2**/ 84.0
Argumentation detection	**78.3**/ 70.2	**79.2**/ 74.4	**82.2**/ 77.3

Table 1: Recognition F-measure of **typical** and tail cases given original and expanded datasets

References

Galitsky, B., Gabor Dobrocsi, Josep Lluis de la Rosa, 2012. Inferring the semantic properties of sentences by mining syntactic parse trees. Data & Knowledge Engineering, V81-82 pp 21-45.

Galitsky, B. 2017. Matching parse thickets for open domain question answering. Data & Knowledge Engineering, v 107, pp. 24-50.

Robinson JA. A machine-oriented logic based on the resolution principle. Journal of the Association for Computing Machinery, 12:23-41, 1965.

Plotkin, GD A note on inductive generalization. In B. Meltzer and D. Michie, editors, Machine Intelligence, volume 5, pages 153-163. Elsevier North-Holland, New York, 1970.

Khardon, Roni and Marta Arias. 2006. The subsumption lattice and query learning. Journal of Computer and System Sciences. v 72, Issue 1, February 2006, pp 72-94.

Mikolov, Tomas, Chen, Kai, Corrado; G.S., Dean; Jeffrey. 2015. Computing numeric representations of words in a high-dimensional space. US Patent 9,037,464, Google, Inc.

Marcu, D. 2000. Rhetorical Parsing of Unrestricted Texts. Computational Linguistics V 2 N3.

Ettinger, Allyson, Sudha Rao, Hal Daumé III, Emily Bender. Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task. 2017. EMNLP, Vancouver, Canada.

Kovalerchuk, B and Kovalerchuk, M. 2017. Toward Virtual Data Scientist with Visual Means. IJCNN.