2 Approach

In this chapter, I will give a brief overview of several things: In a first step, I will elaborate on the topic of pretrained embeddings, transfer learning, and BERT, which is probably the most successful architecture implementing these concepts. Secondly, I will shortly demonstrate problems that have been identified relating to the performance of BERT and its relation to semantics. Follwoing, I will point out some submitted solutions countering those problems. Lastly, I will describe my approach and finally elucidate the topic of semantic roles.

2.1 Pretrained Embeddings and Transfer Learning

In recent years, it has proven to be very effective to use word embeddings precomputed by sentence or document encoders as input representations for task-specific architectures, which may fin-tune those embeddings while learning the task at hand. Pretrained word embeddings have the advantage that they don't need to be learned from scratch by system solving the task at hand.

Bengio et al. [2003] first implemented distributed word representations by using a neural n-gram language model which could be re-uesd in downstream tasks. In "Natural language processing (almost) from scratch", Collobert et al. [2011] showed tHE utility of such embeddings for representing text as input for other neural networks which addressed a multitude of classical NLP tasks. Mikolov et al. [2013] proved that such embeddings could be computed using a modest one-layer neural network targeted at selfsupervised training obejctives, namely CBOW and skip-gram. Context-sensitvity, i.e. taking into account the surrounding words in the embeddings, was introduced by the ELMo architecture Peng et al. [2019] and representing words using pretrained, contextualizing embeddings has since become standard in NLP.

The architecture computing contextualized word embeddings that has caused the most uproar recently was probably BERT Devlin et al. [2018], a model that led to so many variants of it, that it created a whole new field inside the NLP community — winkingly baptized as "BERTology" Rogers et al. [2020]. BERT is a good example for

a typical neural age NLP model: It's architecture is completely agnostic of symbolic, or structural knowledge about language whatsoever, it "only" operates on sequental concatenations of symbols; it also does not employ preprocessing of the data of any kind — no POS-tagging, no dependency parsing, no NER.[1] Albeit this complete lack of any sort of explicit linguistic knowledge, by only extracting statistical patterns which it learns from processing huge amounts of text targeted at its pretraining tasks, the resulting BERT embeddings achieved several SOTAs on well-established NLU data sets, such as GLUE Wang et al. [2018].

The basic concept of BERT is described by the Devlin et al. as being a two-stepped framework: (1) Pretrain the model on unlabeled data over different pretraining tasks[2] and (2) for downstream tasks, use the embeddings by initializing the model with the pretrained parameters and fine-tune all of the parameters by using labeled data from the task at hand.

One of the distinctive features of BERT is the minimal difference in architecture between pretraining and finetuning: The transformer based network which computes the contextualized token emeddings during the language modeling pretraining is also used in downstream tasks only with a so-called task-specific head on top of it. This makes re-implementing the BERT model in other architectures quite convenient, which is one of the reasons I decided to implement BERT in my approach as well.

Another advantage of BERT is that the cost, hardware, and data intensive pre-training of the embeddings must only be computed once; the downstream task dependent fine-tunig can then be carried out in a lean set up.[3]

[1] Because of this non linguistic specific architecture, BERT can easily be adapted to operate on other sequential data. This has actually been done: Ji et al. [2020] for example trained a DNABERT model for successfully deciphering non-coding DNA.

[2] Devlin et al. use the BooksCorpus, consisting of 800 million words, plus the English Wikipedia, consisting of 2.5 billion words. BERT optimizes its parameters on two training objectives: (1) Presented with a sentence containing one random word masked, the model has to predict it, and (2) BERT has to decide if, given two randomly sampled sentences, the second is a valid continuation of the first. Crucially, both tasks can be generated automatically, no tedious human annotation of data is needed.

[3] To give an impression on the expenses of pre-training the BERT architecture: Schwartz et al. [2019] estimate the pre-training for BERT-large to have lasted four days on 64 TPU chips, resulting in power expenditures of about $7,000. However, this has to be considered rather cheap compared to recent architectures' sizes: The largest architecture to this date is the T-NLG (Turing Natural Language Generation) built by Microsoft, possessing a staggering 17 billion parameters — that is approximately 48 times the size of BERT-large (350 million parameters), cf. Sharir et al. [2020]. Open AI's GPT-3's [Brown et al., 2020] pretraining is

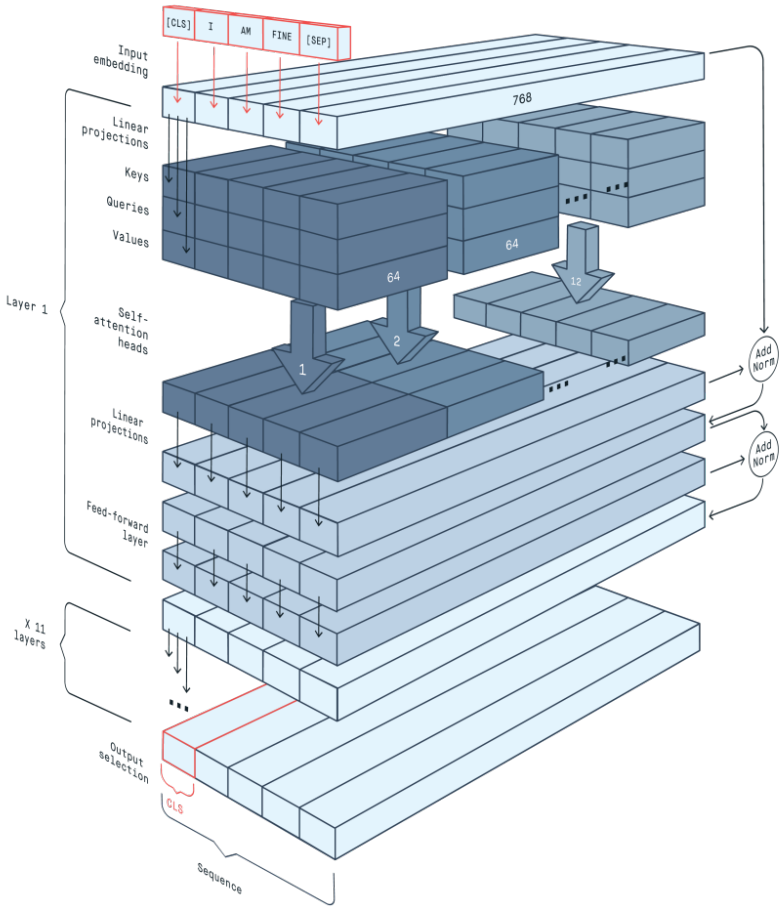

Figure 1: 3D-visualization of the BERT architecture. The 12 attention heads and following linear projections in one block (here called "Layer") are vividly representend. The information flow is top-down, with the tokenized input sentence starting with the [CLS] token on top, and the computed embeddings on bottom. Credit for the image goes to Peltarion

From a more technical point of view, BERT is first and foremost a multi-layered bidirectional transformer encoder. At it's heart lies an implementation of the self-attention mechanism, the transformer, introduced by Vaswani et al. [2017]; the main difference, and apparently advantage, to other architectures that implement the transformer architecture, e.g. Open AI's GPT [Radford et al., 2018], is that BERT is a bi-directional architecture. In simple terms, BERT takes an input sequence, tokenizes it and computes contextualized vector representations for each token via

estimated to have costed $12 million [Floridi and Chiriatti, 2020]. This trend of increasingly bigger language models has earned severe critique from several sides, ranging from ecological and social to linguistic concerns over such models; for a good overview of these points see Bender et al. [2021].

stacked blocks, depicted in figure 1, employing self-attention and linear combination. For reasons of space, I will not go into more detail here about BERT's architecture; the interested reader is referred to the original paper, as well as Clark et al. [2019], which provide an insight into the inner workings of BERT's attention mechanisms.

2.1.1 Problems, Weak Spots, Semanticity

In recent years, a lot of research went into analyzing, improving, and deluding the BERT architecture; In the NLP field, those efforts are now being referred to by the winking notion of "BERTology" (cf. Rogers et al. [2020]).

While BERT showed to be highly effective on several established data sets and benchmarks such as GLUE [Wang et al., 2018], it soon became obvious that it also had its weak-spots: Ettinger confronted BERT with three language tasks originally coming from psycholinguistics which are well-known to be difficult to tackle. They find, that "it struggles with challenging inferences and role-based event prediction—and it shows clear failures with the meaning of negation".

Apparently, while being able to solve NLU tasks to a impressing extent, BERT seems to be prone to fail in situations where proper semantic understanding of text sequences is crucial, such as detecting role reversal or sentence completion tasks.

Jiang and de Marneffe [2019] also state "that despite the high F1 scores, BERT models have systematic error patterns", which for them suggests "that they still do not capture the full complexity of human pragmatic reasoning."

Jin et al. [2020] go even further and creat so-called adversarial attacks for laying open weak spots of BERT: After observing that BERT seems to rely only on the statistical cues of a small number of the input tokens to form its predictions, Jin et al. create a sophisticated algorithm to permute the input sentences wihtout actually changing its meaning. Following is an example from the SNLI [Bowman et al., 2015] dataset, exemplifying one such attack:

(2.1) **Premise**: A child with wet hair is holding a butterfly decorated beach ball.

Original Hypothesis: The *child* is at the *beach*.
Adversarial Hypothesis: The *youngster* is at the *shore*.

The italicized words are the ones that were affected by the adversarial algorithm. Obviously — for a human — the meaning of the adversarial hypothesis has not changed from the original one. One could argue that there is a slight difference in style (the adversarial sounds somewhat overblown to me) but it would still count as

an entailment of the premise. However, as the authors report, BERT is affected by such attacks and changes its predictions.

For SNLI, Jin et al. report to bring down BERT's original accuracy of 89,4% to an astonishing 4,0% by permuting 18,5% of the input tokens. Recall that the permutations are, simply put, nothing else than exchanging words with identical meanings.

2.1.2 Solutions / Related Work

But the NLP community was not only passive and/or destructive concerning BERT, it also produced a vast number of adaptions, variations, and improvements to the "vanilla"-BERT. The motivations behind all those BERTlings are as manifold as one can think: some are simply adaptions to other languages than English, some are general variations (in hope of improvement) of the BERT architecture, other are explicitly addressing above outlined problems.

The following overview sheds some light on the iridescent potpourri of the BERT family.[4] Note however that this is by no means an exhaustive presentation of all BERT-variants produced so far.

Adapting to other languages One of the most straightforward modifications to the BERT model is to pre-train it on different langauges. Examples for this are the French CamemBERT Martin et al. [2019], the Italian AlBERTo Polignano et al. [2019], or the Dutch BERTje de Vries et al. [2019].[5]

Adapting to specialized domains BERT was pre-trained on two corpora: (1) The BooksCorpus [Zhu et al., 2015], consisting of 800 million words and comprising 16 different genres. (2) The English Wikipedia, lists and tables excluded (2.5 billion words). Examples of BERTs pre-trained on specialized domains are for example BioBERT Lee et al. [2020], which is an adaption to biomedical language, and LEGAL-BERT Chalkidis et al. [2020] which is a whole family of BERT models pre-trained on legal texts.

Including different modalities Another, highly interesting amplification of the BERT architecture is the inclusion of additional modalities, for exmple (moving) im-

[4] This compilation is partly drawn from the towards-data-science article "A review of BERT based models" by Ajit Rajasekharan.

[5] Devlin et al. [2018] also trained a multi-lingual BERT (mBERT), which was trained on 104 languages. However, language-specific BERTs have been shown to be more performant than employing the mBERT.

ages: VideoBERT Sun et al. [2019a] learns embeddings for image-enriched texts and can be used for example for image captioning or image classification tasks. Several researchers claim that the future of NLP relies on combining text with sensory, e.g. visual, data for creating more stable and reliable models; cf. [Bisk et al., 2020; Bender et al., 2021].

Optimizing architecture/training objective(s) Several BERT-variations modify the actual architecture of BERT: DistilBERT Sanh et al. [2019] is a variant 60% of the size of the original BERT while retaining 97% of its original performance. RoBERTa Liu et al. [2019] essentially modifies core hyper-parameteres such as batch-size, byte-level BPE, and the like, creating a more stable BERT. DeBERTa He et al. [2020] modifies the attention mechanism and the position encoding, while TransBERT Li et al. [2021] introduces a new pre-training framework.

Incorporating structured information One of the strengths of BERT is the unsupervised pre-training on unstructered, raw text. However, research has shown that including structured linguistic information can stabilize BERT and even counterbalance some of the known weaknesses (see above) to some extent: ERNIE Sun et al. [2019b] includes a knowledge graph into BERT, making structural fact representations available to BERT. VGCN-BERT Lu et al. [2020] combines a Vocabulary Graph Convolutional Network with the standard BERT and Zhang et al. [2019b] include semantic role labels in their SemBERT during fine-tuning.

2.2 GliBERT

Infected by the pandemic BERT-fever and persuaded by the proven transfer-learning capabilities of BERT, I decide to also use BERT as pretrained, contextualized word representations in my architecture; precisely, I concentrate on German and chose the German BERT, pretrained and provided by deepset. Concentrating on the weak spots identified relating to semantic understanding of language, my approach will combine those BERT embeddings with structured linguistic information to counter those shortcomings.

There exist several linguistic structures which could be hypothetically included into BERT: GermaNet [Hamp and Feldweg, 1997] is a large lexical-semantic net that relates noun, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets and by defining semantic relations between these

synsets; it can also be characterized as a thesaurus or a light-weight ontology. One difficulty relating to such structures is that the encoding of such hierarchical information and the sucsequent combining with contigouus data — the sequence of BERT-embedded tokens in a text — is not straight forward and requires an elaborate pipeline of different subsystems.

Another, more plain, possibility would be to implement some linguistic-semantic "mark-up" of the tokens in a text: E.g. Dependency-parse the text and concatenate the BERT-embedded head tokens with a numerically encoded representation of the directed binary grammatical relation that it govern ("direct object", "determiner", etc.).

However, I decide to employ semantic roles for several reasons: Zhang et al. demonstrated successfully the feasibility of this undertaking for English. Further, it occurred to me to be a good balance between two extremes: (1) Including sophisticated knowledge structures, or semantic information, which would require extensive preprocessing (stemming, identifying content words, potential word sense disambiguation, look-up in the knowledge base) and rather cumbersome encoding; and (2) straight forward, on the fly "mark-up" of input text, with low information substance in the case of named entities. with semantic roles, I get the best from both worlds: Relatively easy to implement structured information, as will be seen, while — hopefully — truly adding semantic substance to the vanilla BERT embeddings.

Semantic Roles and their goals, history, and applications are described in more detail in the following section.

Since it is common practice to give your enhanced/variegated BERT architecture an appropriate name I decided to not deviate from this tradition, and decided to call my breed **G**erman **l**inguistic **i**nformed BERT, or short: **GliBERT**.

All code relating to the following dataset set ups, GliBERT architecture, and training can be found in my GitHub repository.

2.3 Semantic Roles

One difficulty a system targeted at NLU must tackle is the ability to cope with the vast amount of flexibility and freedom in natural language to express or describe one and the same state of affairs. There may be subtle differences in emphasis, markedness, or style, but the following sentences all roughly speak about the same

states of affairs:[6]

(2.2) The ship leaked critically due to the big waves and went down.

(2.3) Severely damaged by the hurricane, the vessel sank to the ground.

(2.4) The crew — unable to save the stricken freighter — had to be evacuated by air.

Although these three sentences make use of very different vocabulary — an unweighted BLEU score is virtually zero between them — it is obvious to a speaker of English that they all convey more or less the same meaning, that all of them tell the same state of affairs: A ship sank because of the forces of nature. In linguistics, this is often referred to as proposition: The "lexical kernel of of a sentence that determines its truth conditions, regardless of the syntactic form and lexical filling of the given form of expression" [Bussmann, 2006, p. 959]. In other words, the three sentences above describe the same states of affairs, which means that they are verified or falsified by the same conditions in the real world — i.e. *if* what they are denoting *is* really the case.

Maybe the most obvious way of encoding the same proposition with different words is synonymy: "Ship", "vessel", and "freighter" all refer to the same object in the examples above; having several options when choosing a word to denote something is a paramount feature of human language. As section 2.1.1 showed, simply exchanging a small subset of the words in a sentence by synonyms ot them, has potentially severe impact on models which are aimed at NLU tasks on such sentences. In other words, BERT apparently fails often to recognize that two sightly different word sequences had the same meaning.

Further, the description of one and the same event, e.g. the proposition of the sinking incident of a certain ship, can be linguistically encoded in various ways: In the second sentence, the process is denominated explicitly using the verb "to sink"; in the first, the semantically more obscure semi-fixed expression "to go down" is used to inform about that very situation; while in the third the sinking of the ship is not mentioned explicitly but inferable from the circumstance of "not being able to save" it.

[6] Of course, from an aesthetic, literary point of view, the chose of the right words is crucial and should by no means be played down — "The difference between the almost right word and the right word is really a large matter. 'tis the difference between the lightning bug and the lightning", as Mark Twain famously put it.

For a human speaker, all this disentangling, recognizing coreference, reconstructing not explicitly mentioned information, etc. happens effortless and automatic — for an algorithm, however, phenomena like the ones mentioned before pose serious challenges. In other words, despite the differences that may be encountered on several linguistic levels, as in vocabulary, word ordering, emphasis, etc., there is a remarkable capability in human language processing which reliably extracts the propositional content out of any linguistic statement.

Therefore, a way of equipping an NLU algorithm with tools aimed at semantic interpretation capabilities is a core issue that needs to be addressed: "For computers to make effective use of information encoded in text, it is essential that they be able to detect the events that are being described and the event participants." [Palmer et al., 2010]

As I laid out in section 2.1.1, in modern, purely data-driven models like BERT, all linguistic, semantic, and factual knowledge the model acquires, is infered, or learned by it implicitly from raw text data. Nevertheless, and this is what also what intrigued the NLU field to it, BERT seems to perform surprisingly good in tasks where such "understanding" of events are being tested.[7] Simultaneously, some investigated failures of BERT seem to indicate that this "understanding" goes not too deep and e.g. the recognition of proposition equivalency between sentences is partially poor.

Semantic Roles are an attempt at creating an instrument with which it is possible to analyze the meaning of sentences in a structured manner and being able to express in generalizable terms their semantic properties, e.g. that two sentences express the same proposition. The central idea hereby is that every utterance has an underlying semantic structure[8] (sloppily phrased: "Who did What to Whom, and How, When and Where?") which can be realized in different surface structures. There have been various undertakings in creating a vocabulary for describing such structures, putting the focus on different aspects and showing varying degrees of analytic detail.

The work "The Case for Case" [Fillmore, 1967] is often seen as the starting point for the theory of semantic roles in modern linguistics. Fillmore argued in it that what he called "Deep-Cases" play a crucial role in the Deep-Structure of sentences — the hitherto prevalent view in Generative Grammar was that case was a purely Surface-

[7] Of course, one does not really measure epistemic understanding in such tests, but this is maybe the closest we get (cf. Sahlgren and Carlsson [2021])

[8] Often, especially in Generative Grammar traditions, this level is also known as deep structure, or D-structure.

Structure related phenomenon and only one of several possibilities to realize syntactic relationships. Interestingly, these "Deep-Cases" were semantically-motivated; in combination with so called verb-frames these cases would capture the semantic core of the proposition embedded in the Deep-Structure: A verb like *open* would e.g. form a frame denoting an "opening event" involving an actor, the "Opener", and an object, "the thing opened" (cf. [Fillmore, 1967, p. 46f.]). Seven of such Deep-Cases were proposed by him, e.g. the "Agentive": " [T]he case of the typically animate perceived instigator of the action identified by the verb", or the so-called "Factitive": " [T]he case of the object or being resulting from the action or state identified by the verb, or understood as a part of the meaning of the verb" [Fillmore, 1967, p. 46]. The Deep-Structure of a sentence would then be realized via certain transformational rules as actual, linguistic utterance; e.g. the question "Did he really go to school in XYZ?" and "He went to school in XYZ" are realizations of the same underlying Deep-Structure. Once these transformational rules and Deep-Structures would be understood and described in a sufficient manner, one could turn the analysis around and by examining the observable surface structures of an utterance via the mapping rules to the underlying Deep-Structures would provide an instrument of retrieving the underlying semantic structure, or proposition.

Building upon those core concepts introduced by Fillmore, other linguists added features to the project of formalizing the core semantic structures of propositions, as summarized by Palmer et al.: In the beginnings of the 70ies, Jackendoff [1972] expanded and refined Fillmore's model by introducing the concept of primitive conceptual predicates and their property of governing arguments, which were conceptualized as bearing some proto-semantics, similar to the Fillmorian Deep-Cases. This approach, known as "Lexical Conceptual Structure" (LCS), proved to be an elegant theory and capable of generalizing well between different verbs; in the 90ies LCS was implemented as system for representing semantics in early NLU and translation models [Palmer et al., 2010]. But, due to its detailed analysis of verbs into (several) primitive predicates and the highly verb specific conceptualized semantic roles of them, LCS turned out to be cumbersome to extend to the whole range of a vocabulary of a language.

Dowty [1991] in contrast, approached the problem of constructing a framework for analyzing core conceptual semantic structures from a different angle: Instead of providing a detailed description of the primitive predicate and idiomatic argument structure for each individual verb, he attempted to identify general functions of noun phrases in what he called "thematic proto-roles". To accomplish this, Dowty drew from the theory of "family resemblance" and defined a set of attributes which would indicate such a thematic role. "The hypothesis put forth here about thematic roles

is suggested by the reflection that we may have had a hard time pinning down the traditional role types because role types are simply not discrete categories at all, but rather are cluster concepts [...]" [Dowty, 1991, p. 571]

For example, [Dowty, 1991, p. 572] defines the property bundle of semantic Proto-Agents as follows:

a volitional involvement in the event or state

b sentence (and/or perception)

c causing an event or change of state in another participant

d movement (relative to the position of another participant)

e (exists independently of the event named by the verb)

Similar clusters of characteristics can also be defined for other proto-roles, like patients or themes etc. However, like most theories in linguistics, Semantic Roles remain a disputed topic in the field until today: "There may be general agreement on the cases (or Thematic Roles or Semantic Roles) [...], but there is substantial disagreement on exactly when and where they can be assigned and which [...] should be added, if any" [Palmer et al., 2010]. However, the general, agreed upon objective of semantic roles and similar concepts may be paraphrased as follows:

(2.5) **Semantic Roles are systematic abstractions of semantic functions that are attributed to the participants in a proposition expressed in human language.**

The volitional acting entity in a situation, e.g. is abstracted as "(Proto-)Agent"; regardless of the actual, concrete event denoted by the verb. Similarly, noun phrases which denote participants that undergo some state of change are captured as "proto-patients".

(2.6) He fears bears.

(2.7) His fear of bears.

(2.8) He is afraid of bears.

For some time now there are lexical resources implementing one approach of structural semantic annotation: For example the PropBank [Palmer et al., 2005], which adds "a layer of predicate-argument information, or semantic role labels, to the syntactic structures of the Penn Treebank"; or FrameNet [Baker et al., 1998], which produces "frame-semantic descriptions of several thousand English lexical items and

backing up these descriptions with semantically annotated attestations from contemporary English corpora".

I focus here on the PropBank approach since the semantic role labeler I use in my architecture was trained on the German CoNLL '09 [Hajič et al., 2009] part, which consisted in semantic role labeling according to the PropBank scheme. In PropBank each verb or rather, each verb sense gets a so-called frame attributed: This consists of a definition of the event denoted by that particalur verb sense and the semantic arguments associated with. For example, the first frame of the verb *to sink*[9] is analysed as follows:

sink.01 (cause to) go down, esp into water, downward motion

Roles:

Arg0-PAG causer of sinking (vnrole: 45.4-agent)

Arg1-PPT thing sinking (vnrole: 45.4-patient)

Arg2-EXT EXT

Arg3-DIR start point

Arg4-GOL end point, destination

Arg5-MNR instrument (vnrole: 45.4-instrument)

Applied on the sentence 2.3, this would lead to the following PropBank annotation:

(2.9) Severely damaged [$_{\text{Arg0-PAG}}$ by the hurricane], [$_{\text{Arg1-PPT}}$ the vessel] [$_{\text{Rel}}$ sank] [$_{\text{Arg4-GOL}}$ to the ground].

Although the propbank annotations are verb-specific, there are some generalizations in the first number of arguments as to what proto-role they denote, as the authors of PropBank write: "For a particular verb, Arg0 is generally the argument exhibiting features of a Prototypical Agent [...], while Arg1 is a Prototypical Patient or Theme" [Palmer et al., 2005, p. 75]. In their English PropBank Annotation Guideluines, Bonial et al. [2012] nevertheless propose generalizations even for the higher arguments:

Arg0 agent

[9] `http://verbs.colorado.edu/propbank/framesets-english-aliases/sink.html`

Arg1 patient

Arg2 instrument, benefactive, attribute

Arg3 starting point, benefactive, attribute

Arg4 ending point

ArgM modifier

Rel Relation (can be a verb, noun, or adjective)

That is to say, even if the higher numbered argument's proto role is semantically somewhat fuzzier than for argument zero and one, there are potentially generalizable patterns from which a model which is provided with such annotated sentences might still be able to detect task-supportive structures in the data.

Following are some exmample sentences from the PropBank frames[10]. Semantic roles are highlighted using the colors from the previous list. Note that only one relation is marked in the sentences, even if there are multiple. Since DAMESRL only treats verbs as semantic roles distributing relations, I include only verbal "Rel"s in the examples:[11]

(2.10) [Arg0 Yasser Arafat] has [Rel written] [Arg2 to the chairman of the International Olympic Committee], asking him to back a Palestinian bid to join the committee.

(2.11) Once [Arg0 he] [Rel realized] [Arg1 that Paribas's intentions weren't friendly], he said, but before the bid was launched, he sought approval to boost his Paribas stake above 10%.

(2.12) [Arg1 National Market System volume] [Rel improved] [Arg4 to 94,425,00 shares] [Arg3 from 71.7 million Monday].

(2.13) [Arg0 The new round of bidding] would seem to [Rel complicate] [Arg1 the decision making] [Arg2 for Judge James Yacos].

(2.14) The action followed by one day an Intelogic announcement that it will retain [Arg0 an investment banker] to explore alternatives "to [Rel maximize]

[10] accessible through this GitHub repository

[11] To me, not all annotations in PropBank are beyond all doubt; for example, in sentence 2.14 "an investment banker" is labelled as agent "maximzing" the the patient "shareholder value" — however, I would argue that it's rather the "alternative" that take proto-agentive role in maximizing the shareholder values.

[Arg1 shareholder value]," including the possible sale of the company.

(2.15) [Arg0 He] [ArgM-mod would] scream and [Rel cut] [Arg1 himself] [Arg3 with rocks].

Thanks to lexical resources such as the PropBank, a multitude of models aiming at labeling sentences with semantic roles are now available. For German, there is e.g. DAMESRL [Do et al., 2018], trained on the CoNLL '09 [Hajič et al., 2009] data (which implements PropBank style SRLs). This is also the semantic role labeler I employ in my GliBERT system. Since I treat the semantic role labeler essentially as a blackbox in my architecture and use it an out-of-the-shelf SRL predictor, I will not go into details about the concrete implementation of DAMESERL, as well as the different approaches to automatic semantic role labeling in general — the interested reader is referred to Do et al. [2018] for the former and e.g. Palmer et al. [2010] for the latter.

Following are some examples of DAMESRL-labelled sentences stemming from the GliBERT corpus (see chapter 3). The sentences are represented vertically with the leftmost column being the actual sentence; each column represents one identified verb (B-V) and its predicted semantic roles, labelled using the BIO-schema[12] Every column right to the verticalized sentence represents one argument-predicate structure for this sentence. Note that the different labels are slightly differently labelled than in the PropBank, as listerd before: `V` ("Verb") stands for `Rel` ("relation"), `A`n stands for `Arg`n (both are abbreviations for "argument").

deISEAR

Ich	B-A0	O
fühlte	B-V	O
[MASK]	B-A1	O
,	I-A1	O
als	I-A1	O
ich	I-A1	B-A0
aus	I-A1	O
Versehen	I-A1	O

[12]Introduced by [Ramshaw and Marcus, 1999], the BIO-schema is a established way of adding a label to each token in a sequence, indicating if it belongs to a certain subgroup, or chunk, of the sequence. For example, to mark the prepositional phrase in a syntagma like "He is running from the bear", one would mark the word beginning the PP with **B**, any other words inside the PP with **I**, and all other words outside of it, using **O**: "He[O] is[O] running[O] from[B-PP] the[I-PP] bear[I-PP]".

schlechte	I-A1	B-A1
Milch	I-A1	I-A1
getrunken	I-A1	B-V
habe	I-A1	O

MLQA

Welche	B-A1	B-A2
Positionen	I-A1	I-A2
muss	O	O
man	B-A0	B-A0
erreichen	B-V	O
,	O	O
um	O	O
die	O	B-A1
von	O	I-A1
Kaius	O	I-A1
angeordnete	O	I-A1
Position	O	I-A1
eines	O	I-A1
Läufers	O	O
einzunehmen	O	B-V
?	O	O

XNLI

Es	O	O	O
war	O	O	O
das	O	O	O
Wichtigste	O	O	O
was	B-A1	O	O
wir	B-A0	O	O
sichern	B-V	O	O
wollten	O	O	O
da	O	O	O
es	O	O	O
keine	O	B-A1	O
Möglichkeit	O	I-A1	O
gab	O	B-V	O

eine	O	B-A1	B-A3
20	O	I-A1	I-A3
Megatonnen	O	I-A1	I-A3
-	O	I-A1	I-A3
H	O	I-A1	I-A3
-	O	I-A1	I-A3
Bombe	O	I-A1	I-A3
ab	O	I-A1	O
zu	O	I-A1	B-A5
werfen	O	I-A1	B-V
von	O	I-A1	I-A5
einem	O	I-A1	I-A5
30	O	I-A3	I-A5
,	O	O	O
C124	O	O	O
.	O	O	O