Saarland University

Department of Computational Linguistics

Master's Thesis

Adversarially Learning to Manipulate the Cognitive Load of Sentences

---

Submitted in partial fulfillment of the requirements for the degree

Master of Science in Language Science and Technology.

Author:

Denis Emelin

Matriculation: 2546557

s9deemel@stud.uni-saarland.de

Supervisors:

Prof. Dietrich Klakow

Prof. Vera Demberg

20. October 2017

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Place	*Date*	*Signature*

Abstract

Language comprehension is a cognitive process which has been shown to elicit mental effort from language users. One metric which demonstrably correlates with the processing difficulty of an utterance is its information density (ID). The ability to manipulate the ID of arbitrary sentences is desirable for any autonomous system tasked with communicating spoken or written information to a human audience. In doing so, such systems can adjust the cognitive effort required of the message's recipient in a dynamic manner, thereby accommodating any contextual changes and user-specific requirements. In order to explore the mechanisms which could enable such functionality, the present work frames ID reduction as a monolingual translation task and subsequently proposes a novel neural architecture designed with the goal of learning a translation function between arbitrary high-ID English sentences and their low-ID counterparts. In doing so, the intended goal is the reduction of mental processing costs associated with the system's output relative to its input, as perceived by a human comprehender. The achieved ID attenuation is estimated via the psycholinguistically motivated surprisal metric. To circumvent data limitations, learning takes place in an unsupervised setting by drawing upon the 'generative adversarial networks' training framework. In a series of experiments, the proposed system is ultimately shown to fall short of the intended goal, prompting a comprehensive analysis of factors potentially contributing to the observed performance. Extensions of the system's architecture via the multi-task learning and reinforcement learning paradigms are also considered. Despite the lack of positive results, the evaluated failure cases offer valuable insights which will hopefully benefit future progress in this direction of cross-disciplinary research.

Contents

1 Introduction

Natural language is essential to human communication – it allows us to exchange complex ideas, give shape to abstract thoughts, and express our wants and needs. More increasingly, it also serves as the preferred means of interaction between autonomous systems and their users, with examples ranging from navigational aids to personal assistants. With this development underway, many challenges remain to be addressed so as to endow such systems with the full range of linguistic expression, enabling them to communicate with their human users in a natural, varied, and contextually-aware manner. While the sources of these challenges are manifold, the vast variability and ambiguity of natural language as well as its inherent limitations as a communication channel rank highly among them. Not only does a generated message have to express the intended content correctly, it moreover needs to be conveyed in a manner which allows the intended interlocutor to interpret it without undue effort.

Historically, the former consideration has received greater attention in the field of linguistics, arguably forming the centerpiece of research on semantic and discourse theory. Within psycholinguistics and computational linguistics, on the other hand, the analysis of the exact manner in which meaning is encoded and communicated has provided fruitful ground for empirical studies. Here, numerous investigations have been launched into examining when and why specific linguistic encodings are preferred over alternatives which carry an equivalent meaning. Such decision-making can be observed, for instance, when a speaker elects to use a short pronoun in place of a longer referring expression or to include optional information into an utterance rather than leaving it out. In these endeavors, researchers have frequently relied on information theory as the mathematical framework in which to interpret the collected observations in a systematic and principled way [54]. By doing so, many of the observed selection phenomena have been related to the amount of information contributed by words upon their inclusion into the message context, with the corresponding amount of information conveyed over time commonly referred to as information density (ID) [9].

1.1 Information Density as a Measure of Cognitive Load

One popular way of estimating ID is the surprisal metric which measures the informational content in bits and, by nature of its formulation, assigns greater informational value to less predictable words, as compared to continuations which are more probable given the context provided by the preceding sentence prefix [23]. Experimental studies have lent credence to this metric in the past, by demonstrating that comprehenders expend a grater degree of cognitive effort when processing less predictable linguistic stimuli, which is consistent with their posited informational surplus [11, 55]. It follows, then, that utterances comprised from word units which are, individually, highly predictable given the context afforded by the utterance itself should be comparatively less demanding of a comprehender. Collectively, this body of research has thus established a strong link between the surprisal scores of an utterance's constituents and the expected amount of cognitive resources which has to be allocated for said utterance's interpretation , i.e. its associated cognitive load.

Applied to language production by autonomous systems, this insight has a number of interesting implications, many of which can be placed under a common umbrella. Specifically, given that human language users adjust the manner in which the informational content they wish to convey is communicated and, in doing so, vary how cognitively demanding the resultant message is likely to be for the intended interlocutor, it would be both desirable and useful for any artificial system which utilizes natural language as a communication channel to be equipped with the same capacity. This is especially important to consider whenever efficient and effective communication with the user is one of the main goals underlying a system's design. Targeted ID reduction, especially, has a variety of potential applications, as in real-life scenarios users' attention to a system's output is rarely undivided. A person engaged in a primary task will only be able to allocate limited cognitive resources toward the interpretation of the system's output, potentially leading to delayed or incomplete comprehension of the communicated content. This, in turn, can have detrimental and potentially dangerous consequences, should the system be depended upon for instruction or expected to provide urgent information.

One possible scenario illustrating this point is that of a car driver focused on the road ahead and relying upon the navigational system to provide auditory directions to the selected destination. The extra effort incurred by the decoding of a cognitively taxing set of directions can be expected to increase the chance of the driver missing a turn should they fail to parse the communicated message correctly within the allotted time frame. Similarly, hard to process instructions may distract the driver from the road, thus increasing the chance of an accident occurring. Other scenarios which illustrate the need for ID reduction engines include urban search and rescue operatives receiving natural language reports from mobile platforms as they scout disaster areas – a highly exhausting and stimuli-rich working environment [53], or users following a complicated cooking recipe as dictated by a personal assistant. To propose one possible answer to this demand is the first main objective of the present work.

An intuitive way of approaching the task of ID reduction is to frame it as a monolingual translation problem. That is, given a source sentence assigned a certain ID value, for instance expressed as the mean surprisal of its constituent words, the goal is to generate a second sentence which preserves the semantic content of the source while simultaneously reducing its ID value. When embarking on this route, several recent developments have to be considered. First, the field of machine translation has undergone a paradigm shift from n-gram based translation models towards the use of deep artificial networks over the course of the last decade, which continue to set standards in translation quality to this day [3]. Second, machine translation modeling almost always presupposes the availability of aligned corpora from which the mappings between the source and target domains (i.e. languages) can be learned. To the author's knowledge, such corpora – at least of sufficiently large size – presently do not exist for the ID reduction task. In order to nonetheless leverage the 'unreasonable effectiveness' [1] of neural nets for this research direction, it is, therefore, not sufficient to adopt contemporary state-of-the-art translation models wholesale, but necessary to modify and extend them so as to account for the unique challenges and constraints presented by the examined problem. In order to solve the primary challenge of the unavailability of aligned ID-variant corpora, the present work draws inspiration from the visual processing community where learning from unaligned resources has been investigated in the context of generative adversarial nets (GAN) [20]

[1] karpathy.github.io/2015/05/21/rnn-effectiveness

1.2 ID Manipulation via Generative Adversarial Nets

Rooted in game theory, the GAN framework allows generative models to be trained in unsupervised manner. This is achieved through an incremental reduction of the distance between an arbitrary source and target data distribution, over the course of the training process. Importantly, the training procedure does not require for training samples originating from either distribution to be aligned, instead only presupposing the existence of two labeled sets of data, each representing a distinct set of entities. Throughout the GAN-assisted training, the trained model is encouraged to shift the source distribution parameters so as to closely resemble those of the target, with the difference estimated by a binary classifier receiving samples from either distribution as its only input during each training step. As these samples can be assumed to empirically approximate the true distribution form which they are drawn, the desired outcome of the training procedure is for source samples to become sufficiently indistinguishable from the target samples, as evaluated by the classification component.

Within the context of ID reduction as it is conceptualized in the present work, the source distribution is represented by a corpus of sentences assigned a surprisal score above a certain threshold, whereas the target distribution is comprised of thematically similar sentences with surprisal scores residing below said threshold. As GANs presuppose data used in training and inference to be continuous, each sentence is represented by a single dense, low-dimensional vector capturing its semantic and syntactic content, rather than by a sequence of discrete word tokens. Henceforth, the former corpus type will be referred to as 'high-ID', whereas the latter is designated 'low-ID', as convenient short-hands. The second main objective of this work is, therefore, to conceptualize and implement an ID reduction engine built around the GAN framework.

In order to achieve this goal, a modular and extendable system is proposed which consolidates within itself several discrete components, all of which contribute jointly towards the intended goal of learning and performing the translation function between arbitrary input sentences and their low-ID counterparts. As the system relies on GANs to achieve its intended goal of information density attenuation, it is henceforth referred to IDGAN, which stands for 'Information Density Generative Adversarial Net'.

The core components of the system are two encoder-decoder networks, initialized as sequence autoencoders – one tasked with learning to translate high-ID sentences into their low-ID counterparts and one with providing ground-truth low-ID sentence encodings, which comprise the target distribution within the adversarial learning setup. Additionally, the IDGAN architecture incorporates a binary discriminator tasked with correctly classifying the two encoding types and a language model used to monitor the achieved ID reduction as denoted by the difference in mean sentence surprisal between the system's input and output sequences. By utilizing each of these components, the adversarial training objective encourages the system to transform arbitrary high-ID sentence encodings in a manner which makes them indistinguishable from arbitrary low-ID encodings to the discriminator. The mapping function through which this transformation is accomplished is learned by the encoder half of the encoder-decoder network receiving high-ID sentences as its input. This encoder network therefore simultaneously functions as the GAN generator. The discriminator, meanwhile, is trained to accurately distinguish between the so transformed, originally high-ID, sentence encodings and true low-ID sentence encodings. The latter are constructed by the second encoder-decoder on the basis of sentences drawn from the low-ID corpus. Throughout the training, the discriminator aims to minimize its classification error, thus learning to be more successful at keeping both categories apart. Conversely, the generator aims to maximize the discriminator's error, in the process learning to encode high-ID sentences into representations containing feature distributions which closely resemble those characteristic of true low-ID sentence representations.

The encodings produced by the generator on the basis of high-ID sentences and made to resemble low-ID encodings as a consequence of the adversarial objective are subsequently decoded back into sentences by the decoder connected to the generator within the shared encoder-decoder architecture. Intuitively, matching high-ID encoding features to those characteristic of low-ID encodings is expected to lower the mean ID of the decoded sentences, relative to their high-ID source. This desired outcome is further encouraged by controlling for sources of high-level linguistic variation other than ID when constructing the ID-variant corpora, as is done as part of this study. The corresponding corpus construction methodology is detailed in chapter 3.1. During the decoding, a reconstruction criterion identical to that employed in sequence autoencoders (SAEs) is relied upon to enforce that the generator's input and the decoder's output be comparable with regard

to their communicated content. Given the non-aligned nature of the training samples originating from either ID-specific domain, content preservation could otherwise not be inherently guaranteed. Lastly, a multi-stage pre-training scheme is employed to facilitate the construction of sentence encodings which are capable of capturing the syntactic and semantic properties of sentences they are derived from to a satisfactory degree. A more detailed, illustrative overview of the system architecture is given in 3.2.

1.3 Thesis Contributions and Structure

Taken in its entirety, the present work seeks to contribute to the ongoing multi-disciplinary research in the field of computational linguistics by:

1. Offering a thorough exploration of the unsupervised ID reduction task
2. Applying adversarial learning to a psycholinguistically motivated problem
3. Defining and evaluating a construction strategy for unpaired ID-specific corpora
4. Providing an expandable, low-level IDGAN implementation adaptable to related natural language processing tasks
5. Conducting an error analysis for several failure cases exhibited by the proposed system in a series of experiments

Following a comprehensive overview which dissects the complementary influences that motivate this work, a detailed description of the proposed ID reduction engine is given, accompanied by an in-depth discussion of data procurement, training strategies, and component evaluations, as well as the difficulties encountered along the way. This is followed by an informed analysis of several unsuccessful experiments during which the IDGAN system had been applied to the task of ID-reduction under different hyper-parameter configurations.

On the basis of this analysis, suggestions are made as to how the system can be further improved and which considerations should guide future investigations into unsupervised ID modulation. A concrete proposal presently under works for the next iteration of the system developed here, which improves upon the current design by learning from the experimental insights, closes of the manuscript.

2 Literature Review

2.1 Surprisal as a Metric of Information Density

For any inquiry into the the cognitive effort associated with the processing of natural language to be scientifically rigorous, the use of a clearly defined complexity metric is paramount. The chosen metric has to offer a sufficient degree of granularity, behave consistently when measured using established methods, and remain firmly rooted in the observable reality of the human mind. Over the years, the linguistic tradition has brought forward several measures satisfying these criteria, with surprisal being one that is frequently referenced in the fields of psycholinguistics and computational linguistics. This is undoubtedly in part due to the relative ease with which it can be approximated by computational models of language, but also due to its strong, empirically established correlation with processing difficulties observed in human language comprehenders. Based on principles of information theory [54], surprisal has been historically applied to estimating the processing difficulty language users experience when perceiving some linguistic expression, typically represented as a sentence string at different stages of completeness.

In its essence, the surprisal score seeks to estimate the amount of information conveyed by a natural language expression, by relating this quantity to the predictability of the word within the context in which it occurs, typically limited to the preceding sentence prefix. As applied to language processing, the intuition is that words which are more predicable, and therefore carry less novel information, require less cognitive effort on the part of the comprehender to be understood. Highly informative words, on the other hand, update the preceding context to a large degree and therefore require a greater amount of cognitive resources to be processed. Thus, upon encountering a hitherto unexpected word in the course of an interaction based in language, the recipient of a verbal message needs to expend more time and effort, as compared to cases where the encountered word is highly probable given the linguistic context perceived by the language user up to this point in time. Within the theory of surprisal, the unexpectedness of a word given some previous context is therefore directly proportional to the amount of cognitive effort required to comprehend it.

Consequently, the calculation of the surprisal score associated with some target word presupposes the availability of a means for estimating the probability with which said word may occur given the preceding context. Probabilistic grammars as well as language models offer a convenient means for obtaining such conditional probabilities, as by assuming a probabilistic perspective sentences can be framed as sequences of discrete word tokens, each emitted at the corresponding time-step with a probability conditioned on the preceding sentence prefix. Given access to such point-wise probability estimates, surprisal can then be calculated according to the equation given in (1), outlined in [23] among others.

$$log_2 \left(\frac{1}{P(y)} \right) \tag{1}$$

Following this definition, information can thus be equated to the log of the reciprocal of an event's probability and ID to the amount of information carried by the individual units comprising a communicated message.

One practical application of surprisal theory which draws upon a computational model of linguistic knowledge, albeit one significantly constrained in scope and domain, can be found in [22]. In this seminal work, Hale combines the Earley parser with a probabilistic context-free grammar to estimate per-word surprisal scores at any point during the parsing process, the totality of which is taken to represent the cognitive load experienced by human subjects. In doing so, a strong competence of grammar, the human parser's sensitivity to frequency effects, and its eager nature are assumed. The observed surprisal effects are argued to arise as the immediate result of the parser discarding the total probability mass associated with partial syntactic structures which had been consistent with the preceding sentence prefix but were found to be incompatible with the novel syntactic information introduced by newly encountered words at each time step over the course of sentence processing. Although the syntax-centric perspective taken by Hale in postulating the specific causes underlying surprisal variation diverges from the aforementioned information-theoretical point of view, both concur with regard to the core idea of word predictability in context – as denoted by the corresponding surprisal rating – having a significant impact on the ease with which sentences are processed by language users. This

shared assumption is therefore also adopted by the present work.

Due to its popularity within the field of psycholinguistics, there exists a substantial collection of empirical evidence in support of the surprisal theory. In extant studies, surprisal effects have been frequently observed to positively correlate with well-established behavioral indicators of processing difficulty and increased cognitive load, such as increased reading times [11], the N400 component in event-related potentials[14], and pertinent brain area activations in fMRI imaging [24].

Evidence for the validity of surprisal as a processing difficulty metric also can be found in language evolution studies. In a series of large-scale computational simulations presented in [17], word orders of five natural languages were evaluated against random pseudo-grammars and grammars which explicitly optimize processing efficiency. Findings from this comparison suggest that naturally evolved grammars exhibit a pronounced preference towards ease of processing as estimated, in part, via surprisal ratings. Its capacity for approximating one source of evolutionary pressure in language evolution further strengthens the ties of the metric to the cognitive reality of language users.

Syntactic reduction – a linguistic phenomenon where speakers choose to forgo the the inclusion of optional word tokens such as relativizers into the produced phrase or sentence – has also been related to surprisal effects. Using a regression model trained on surprisal estimates derived from a language production model, [31] were able to reproduce syntactic reduction variation evident in natural language to a large degree. From the observed model behavior, they conclude that language users exhibit a preference towards realizing optional elements in contexts which are less predictable, i.e. more information-dense, and that both phrasal and structural information is used in the estimation of predictability levels. Obtained through rigorous empirical experimentation, these results, too, provide compelling evidence in support of the metric's significance as one that influences language processing and use in humans.

Although high surprisal scores are not exclusively tied to select sentence types, several linguistic constructions have been identified that reliably elicit significant surprisal effects. A prime example among this category are garden-path sentences [38], such as the famous

example *The horse raced past the barn fell.* It should, however, not go unmentioned that surprisal does not always succeed in predicting processing difficulties observed in human comprehenders. One case for which this holds true are object-extracted relative clauses in the English language. While surprisal estimates correctly predict an increase in reading times for this clause type as compared to the more frequent subject-extracted relative clauses, the location at which the slowdown is expected to occur is different from the actual difficulty locus, as documented in a reading-time study [37].

A different estimate of processing difficulty, the entropy reduction metric, provides the correct forecast in the aforementioned scenario and thus appears to complement surprisal well. However, just like surprisal, it does not make universally valid predictions [23]. Furthermore, approximating entropy reduction computationally is far from trivial, as its calculation requires for all possible sentence continuations to be considered, starting from the time-step at which the measurement is undertaken. This is computationally intractable and even approximate estimations [13] are expensive for sufficiently large vocabularies. While exploratory experiments into the joint usage of surprisal and entropy reduction have been carried out in the initial stages of this project, the approach has been found to be prohibitively costly with respect to time and resources, even by deep learning standards. Surprisal, on the other hand, requires only point-wise word probability estimates, which can be obtained quickly and efficiently. For this reason and in light of the considerable amount of supportive experimental evidence, the current work relies on surprisal as the sole means by which processing costs associated with natural language sequences are measured. Nonetheless, the inclusion of entropy reduction as a metric of interest for the ID reduction task will likely prove valuable for any related research endeavors to be undertaken in the future.

Given their state-of-the-art accuracy, language models based on recurrent neural networks (RNN-LMs) are a compelling choice as the source of word probabilities for surprisal calculation, as has been previously demonstrated by [13]. Not only are RNN-LMs well researched and have found a plethora of application in NLP literature in recent years, their use is also not predicated upon commitment to one specific grammar formalism, which allows for a theory-agnostic, versatile, and easily reproducible surprisal estimation. Applying the surprisal formula to RNN-LM's predictions is straight-forward and can be

accomplished by extracting the probability estimates from the LM's final layer, which traditionally applies the softmax function to the (possibly transformed) output of the recurrent network.

Viewed collectively, the body of evidence for the validity of surprisal as an estimator of language processing difficulty briefly referenced here unequivocally supports the choice of the metric as the primary means of assessing ID within the present work. Its inherent compatibility with neural language models, too, is highly desirable, as it allows for a transparent and easy to interpret integration of surprisal estimates into the training pipeline, as a means to monitor the proposed system's performance. In the next two sections of this review, attention is given to the system's individual components, the chosen learning framework, and the exact means by which these two factors are leveraged to enable automated ID reduction.

2.2 Neural Networks for Language Modeling and Manipulation

Within NLP literature exists a rich history of exploiting the expressive power of neural networks for the task of language modeling. Among the initial successful forays into this research direction is [5], where an LM based on a feed-forward network (FFN) is successfully applied to several large corpora, resulting in substantial improvements in perplexity scores compared to baseline models relying on n-gram information. Since RNNs excel at capturing long-distance dependencies characteristic of natural language sequences within an evolving cell state, especially once equipped with Long Short-Term Memory (LSTM) [27] or Gated Recurrent Unit (GRU) [8] cells, they quickly became the preferred method for implementing neural LMs. The research area of neural language modeling continues to be an active one, with recent studies focusing on investigating its limits, by training refined LM architectures on corpora of exceptional scale [34] as well as by seeking to consolidate RNNs with FFNs for language modeling, so as to combine their respective advantages [44].

As the IDGAN-internal LM is used to estimate surprisal scores, the architectural considerations are guided by efficiency and efficacy factors, rather than the intent to improve upon leading models in terms of perplexity scores. A stacked LSTM-RNN with tied input and output embedding projections fulfills these demands to a sufficient degree, as the enforced parameter sharing reduces the model size while simultaneously improving the model performance [46]. Moreover, such weight-tying has been empirically shown to result in higher quality word embeddings, which further benefits the proposed IDGAN design, as detailed in section 3.2.

For the goal of manipulating the informational content of a sentence, which is determined by factors such as its semantic and syntactic dimensions, the availability of meaningful and, ideally, learnable sentence representations is essential. Obtaining such encodings has been the focus of many recent endeavors in NLP research, with a great number of them leveraging deep learning methods centered around sequence-to-sequence models. The exact nature of the mechanisms used to obtain sentential encodings may differ from one approach to another, and can include attention [59], convolution operations [35], or recurrent and recursive composition [30]. One approach which falls into the latter category

and has found popularity in NLP, is the sequence autoencoder (SAE), initially introduced in [57] and developed further in [10], among others.

Offering a stable training performance and impressive results even for comparatively small datasets, SAEs traditionally consist of two principle components, each realized separately by a neural network – an encoder and a decoder, which may or may not share some of their parameters. The processing of sequential information within an SAE is facilitated through the inclusion of recurrent neural networks (RNNs) at both positions, making this architecture especially well-suited for the processing and analysis of time-sequence data, such as word strings of arbitrary lengths. During the forward pass, the interplay between the two halves of an SAE takes place in two stages. In the initial stage, the encoder receives its input – a sequence of either dense word embeddings or one-hot word vectors – and compresses it into a single-vector meaning representation through a series of affine transformations and applications of non-linear functions, which can then be accessed via the encoder's final hidden layer. In the second stage, the so obtained sentence encoding is forwarded to the decoder net, which is trained to recreate the input sequence by minimizing the observed reconstruction error. Upon the completion of the training procedure, the latent sentence representations generated by the encoder network can be extracted and re-purposed for a number of downstream tasks, where they are commonly taken to represent the compressed content of the original input. While a lower sentence encoding dimensionality offers higher computational efficiency, an increase in dimensions allows the encoder to compress sequences of greater length more accurately [10].

As an inevitable side-effect of the compression procedure, parts of information present in the input sequence may not be carried over into the associated sentence representation, resulting in reconstruction errors. Since the dropped information may be of crucial importance to the goal pursued by the system utilizing the encodings, methods have been developed which enable encoder-decoder models, of which SAEs are one specific instance, to learn which information within the input sequences should be paid attention to. Such attention mechanisms come in various shapes and differ in the particulars of their intended contributions, but all provide models incorporating them with an additional channel through which information that is relevant to the learned task can be propagated. One field which has come to rely extensively on attention mechanisms is neural

machine translation (NMT), where their inclusion has yielded substantial improvements in transition quality for numerous language pairs, as measured by BLEU and other scoring schemes [45]. Among the attention strategies used in NMT, the methods first introduced in [39], referred as 'Luong' attention from here on for convenience, have enjoyed great popularity in recent publications, e.g. [52].

The present work incorporates the global variant of 'Luong' attention combined with input-feeding within both of its component SAEs, with the aim of improving the quality of the sentences generated by the respective decoders. As opposed to its local counterpart, global attention permits the decoder to examine the entirety of the input sequence during each decoding step. While this is computationally more expensive for longer sequences such as paragraphs and documents, global attention performs well for isolated sentences while introducing less additional parameters into the final model than the local variant. Furthermore, the global approach may be better-suited for the ID-reduction task if the attention parameters are jointly learned with the ID-reduction function, as the intention here is to lower the surprisal estimate for the entirety of the input sentence. This, however, is not done in the implemented system due to limitations of the adversarial learning objective discussed in section 3.2.

While a detailed overview of the global 'Luong' attention goes beyond the scope of a literature review, its functionality can be summarized as a sequence of three steps taking place during the decoding phase. In the first step, the attention mechanism takes as its input the hidden states generated by the encoder RNN at each time-step throughout the encoding process. These are subsequently used to derive a context vector which weights the individual constituents of the input sequence according to their relative importance for the decoding task. Lastly, the context vector is combined with the current decoder hidden state and fed through a softmax layer to obtain the predictive distribution for the current time-step. In this manner, an improved correlation between the input and output sequences is achieved.

Word embeddings and sentence encodings, taken together, form the input-side foundation upon which the desired functionality of the IDGAN sysem is predicated. The methods outlined within the previous paragraphs – neural LMs, sequence-to-sequence modeling,

and attention mechanisms – jointly ensure that this foundation is a solid one, obtained in a well-defined, structured, and computationally feasible manner. As the same methodology can be applied to obtain low-CL and high-CL encodings alike, the information-theoretical spectrum of data to which the ID reduction engine is given access is therefore fully attended to. However, as the goal of the current research endeavor does not merely lie in the generation of sentence representations but in their modification through distributional alignment, the adversarial learning mechanism by which this is to be accomplished, too, requires a closer examination. This is the purpose of the next section.

2.3 Distributional Alignment through Adversarial Learning

With the procurement of word- and sentence-level information well-suited for a task which involves the joint manipulation of sentence meaning and structure handled, the exact process by which the goal of ID manipulation can be achieved can take on multiple forms. The breath of available options shrinks rapidly, however, once the decision is made to eschew the time- and resource-intensive collection of aligned corpus data corresponding to the ID levels of interest. Given this constraint, unsupervised learning appears to be the obvious direction to consider, as the lack of explicit targets for the considered training samples is a characteristic starting point for this approach. The task under scrutiny here is atypical for the unsupervised learning domain, however, resembling on its surface a supervised translation task in which one sentence is transformed into another, so that the communicated content is preserved while some sentence modality – either the language identity or, in this case, the sentence-specific surprisal value – is altered.

Approached from the traditional machine translation perspective, the mapping function between the source and target domains can be trivially learned by a sequence-to-sequence model aligning individual sentences from one corpus to their corresponding counterparts in the other. As this is not a possibility in the scenario considered by the present work, due to the absence of sentence-aligned ID-variant corpora, other means by which this goal may be achieved must be considered.

One such alternative is to learn the mapping function not from individual examples, but by trying to align the source and target distributions in their entirety, altering the source distribution in a way that makes samples drawn from it indistinguishable from samples originating from within the target distribution. The ideal outcome, then, is that a model trained to perform this alignment will learn to map any point within the source distribution, as specified by the input data it receives, to the corresponding point within the target distribution, as determined by the learned alignment. Importantly, this mapping should ideally preserve domain-invariant qualities of the input data. This is core conceit behind the GAN training procedure. Although GANs have come into the public spotlight only recently, the body of work built around the framework has been rapidly growing since its initial conception, with publications proposing alternative adversarial

learning objectives [41, 62], modifying the traditional relationship between the generator and discriminator networks [12], and applying GANs to problems outside of the visual domain where they originated [26]. As part of the latter category, recent forays into GAN-based language generation have shown interesting and promising results [60].

Irrespective of their specific application, models utilizing GANs commonly bear similarities in their base design to the seminal contribution of Goodfellow et al., in which the adversarial objective function was first formalized and consequently applied to the generation of natural images from noise samples. The often-cited metaphor capturing the core functionality of GANs and offering a high-level, immediate means of understanding the individual contributions of its two components, is that of a competition between two adversaries – a group of counterfeiters and the police. The counterfeiters' goal is to manufacture fake currency which is indistinguishable from that in circulation, while the police must distinguish the fake banknotes from the real ones. Within the GAN paradigm, the counterfeiting is done by the generator model, typically implemented as a task-specific neural network, while the policing action is performed by the discriminator, a binary classifier traditionally given form of a FFN or convolutional neural network (CNN). Both GAN components are trained jointly in a min-max game that converges as soon as a Nash equilibrium is reached, which is the case once the performance of either network can no longer improve under the assumption that its adversary's behavior remains consistent. A formulation of the adversarial objective function is given by equation (2).

$$\min_G \max_D V(D,G) =_{x\sim p_{data}(x)} [logD(x)] +_{z\sim p_z(z)} [log(1 - D(G(z)))] \tag{2}$$

Here, G denotes the generator, D is the discriminator, x represents a sample from the target data distribution, while z is a vector of continuous values drawn from the source distributions (e.g. of Gaussian noise). As the training goes on, G is trained to transform any input data in a way that renders it indistinguishable to D from any target sample. To accomplish this, G is trained using the classification error gradient of D. As D learns to perform better on the classification task, its decisions form a point of reference for the adjustment of G's parameters via backpropagation. By adjusting G's parameters in the

opposite direction of D's gradient, G's likelihood of generating fake data which would be classified as belonging to the target distribution by D increases. Over the course of many iterations of this process, G ends up learning a mapping function by which the source distribution is moved towards the target distribution until the two are aligned. Alternatively, the GAN objective as formalized in (2) can be described as an approximate minimization of the Jensen-Shannon divergence between the two distributions, for which a formal proof can be found in [20]. Importantly, GAN-assisted training requires for the training materials to be continuous in nature, which is due to multiple reasons. On the one hand, the incremental adjustments to the statistical properties of the source distribution taking place throughout the adversarial training are of small magnitude and, as such, insufficient to effect meaningful changes to discrete values [19]. Moreover, it assures the unhindered backpropagation of the adversarial loss signal, which would otherwise be interrupted by discrete operations such as the sampling of word tokens form a predictive distribution.

Among the ranks of the numerous applications of deep adversarial learning that have emerged in the years following its popularization, a not insignificant subset has sought to combine the unique capabilities offered by GANs for distribution manipulation with the relative stability characterizing autoencoders. This decision is motivated by several considerations, such as the desire to stabilize the notoriously fickle adversarial training process and to extend the generative capabilities of sequence-to-sequence models. The adversarial autoencoder (AAE) defined by Makhzani et al. [40] is one representative of this group, bringing together both components within a unified system designed to perform variational inference. It accomplishes this by establishing a link between a source data distribution and some imposed prior – i.e. the target distribution – via the hidden code generated by the autoencoder's encoder network.

Over the course of AAE's training, the GAN discriminator is provided with latent representations of the input data extracted from the encoder as well as samples drawn from the prior, and is trained to distinguish between the two classes. This, in turn, forces the encoder, which simultaneously fulfills the role of the GAN generator, to find a mapping between the source and target distributions, so as to maximize the discriminator's error and reach the GAN-game equilibrium. The autoencoder's decoder, meanwhile, learns

to recreate the original input data from the so transformed hidden representations. A reconstruction error penalty guides the decoder in its training and combines with the adversarial training cost into a compound objective function. Upon the convergence of the training procedure the decoder has thus learned a second generative model, which maps the imposed prior to the data distribution, serving as a mirror image of the encoder model.

The AAE model, on the whole, bears a strong similarity to the envisaged IDGAN architecture, which must be acknowledged at this point. However, each system has been designed with a different goal in mind – applications to the abstract and visual domains at various levels of supervision in case of the AAE and unsupervised ID attenuation in natural language sequences, in case of the current proposal. Additionally, the exact architectural features are not fully identical between the two systems, as described in section 3.2. Nonetheless, while the author was unaware of the AAE publication at the time of the initial conceptualization of the ID-reduction engine, it has motivated several of the subsequent design decisions.

While the AAE bears clear relevance to the goals set forth by the present investigation, its experimental applications remain confined to abstract high-dimensional data and the visual domain. The initial omission of NLP as a testing ground for AAEs has since been attended to by several research projects, seeking to leverage their potential for tasks encompassing NMT, representation learning, and stochastic sentence generation. The first one to find mention here is Barone's foray into utilizing AAEs for the adversarial learning of a mapping between word distributions of two distinct languages, English and Italian, in effect putting forth a method for limited unsupervised bilingual NMT [4]. Although not yielding competitive results, Barone's work nonetheless succeeds in demonstrating the capacity of adversarially trained models for the transfer of semantic information between representations that define distinct data distributions. One example supporting this is the learned connection between the English term *comics* and such Italian expressions as *episodio* and *romanzato*, both carrying a relation to the comic domain.

Another notable example of adversarial solutions to open challenges in NLP can be found in [61]. Here, the aim is the generation of realistic, natural sounding, text on the basis

of latent representations drawn from a noise distribution. The learned functionality can therefore be compared to unsupervised language modeling. The adversarial training is relied upon for enforcing a smoothness of the learned distribution and for enabling the sampling of meaningful, well-formed sentences from anywhere in the posterior. This is achieved by feeding the discriminator with sentences generated by the generator from the latent code and natural language sentences drawn from the target corpus, while the discretization problem is avoided by obtaining the training signal via the REINFORCE algorithm with Monte Carlo sampling, as opposed to e.g. the discrimination between continuous sentence encodings. The result of this design is an AAE capable of generating a continuous latent representation space that encodes plausible sentences at any of its coordinates, from which a sufficiently accurate reconstruction of the source domain samples is shown to be possible.

One last example for the transfer of GAN assisted learning into the natural language domain which will be mentioned here is found in [47]. Just as the work of Zhang et al., the goal of this study is to enable the generation of well-formed natural language sequences from noise vectors, rather than attempting to learn a mapping function between two discrete domains. To achieve this, while also resolving the incongruity between the requirement for continuous training data and the discrete character of natural language sentences, the discriminator network is presented with sequences of probability distributions over all vocabulary entries obtained at each time-step from the generator model, rather than discrete word tokens. On the other side, the discriminator also receives series of one-hot word vectors from a target corpus, which correspond to samples approximating the target domain. By training a generative model within the so modified GAN setup the authors report initial results on tasks such as the generation of Chinese poetry and sentiment-conditioned utterances. However, the generated sequences retain a distinctly non-natural quality, as evidenced by the published output samples. As such, the study is offers insights as to the potential GANs carry for NLG applications as well as the need for further development required to make this approach to the unconstrained generation of natural language a competitive one.

The greater picture emerging from the recent developments in NLP research discussed here suggests a significant promise held by adversarial learning for many open tasks involving

natural language generation. This seems more so true of methods which seek to combine GANs with autoencoders for added guidance and training stability. Furthermore, their potential comes to the fore clearly when applied to tasks involving the alignment of arbitrary distributions with the goal of obtaining varied and natural-sounding generation results. Coupled with the information theoretical concepts previously introduced, these observations form the foundation upon which the contribution of the present work is built. How exactly these constituent parts may be effectively combined into an ID-reduction engine and why the resultant system, as formalized and implemented in the context of the present study, ultimately falls short of offering a compelling solution to its assigned task is the subject of the subsequent chapters. First, however, the gathering and processing of the data used in the training and evaluation of the IDGAN is addressed.

3 IDGAN

3.1 Obtaining ID-Specific Data

For the procurement of data on which the proposed ID reduction engine is trained and evaluated, the primary consideration has been to emphasize the divergence in surprisal scores between sentences comprising the low-ID corpus on the one hand and the high-ID corpus on the other. Furthermore, this difference needs to be statistically significant with a sufficiently large effect size, so as to serve as the primary signal for the adversarial training of the evaluated system, enabling it to learn the mapping from arbitrary high surprisal sentences to corresponding low surprisal translations in an unsupervised fashion. To accomplish this, the ID-variant corpora have been constructed in a manner which seeks to control for other, potentially interfering sources of large-scale linguistic variation, such as the genre, domain, and register of the considered source texts.

With these considerations in mind, a large monolingual, single-domain document collection was deemed the ideal starting point for the construction of the ID-specific corpora. The monolingual English Europarl corpus, release v7[2], which is comprised of parliamentary speeches on topics related to European legislation, fulfills these criteria to a reasonable degree. In addition, as it has been frequently drawn upon for the training and evaluation of deep learning models, it represents a linguistic resource that is much studied, largely free of noise, and demonstrably compatible with neural network systems such as the one proposed herein. Existing models trained in part on the English subsection of the Europarl corpus and applied to tasks related to ID reduction, including language modeling and machine translation, can furthermore serve as useful points of comparison for the proposed system's performance. Lastly, its unhindered accessibility removes a potential hurdle for the reproduction of experiments discussed in the subsequent sections.

As the present work is exploratory in nature and aims to evaluate the applicability of the GAN training regime to the ID reduction task by examining various training conditions, and given that training neural networks remains a computationally expensive endeavor,

[2] statmt.org/europarl

only the initial 100k sentences of the overall English Europarl corpus are considered at the outset of the data preparation process [3]. The so truncated data source is relatively small by the standards of the field – by comparison, the Penn Tree Bank corpus is roughly twice as large. However, its limited size is advantageous in that it cuts down the expected training duration significantly, especially for a target system with a high number of parameters subject to optimization as is the case for the full IDGAN setup. This, in turn, allows for a rapid evaluation of different training configurations.

While the limited amount of training data may quickly lead to a model overfitting the training set, regularization techniques including L2-regularization and dropout [51] can be employed to alleviate this concern without noticeably increasing the training speed. Similarly, by stopping the training process as soon as no further improvement of the system's performance on a validation set can be observed for some specified number of training steps – a strategy commonly refered to as 'early stopping' – overfitting is reduced while generalization to the validation set is encouraged. Also worth mentioning is that since the Europarl corpus consists of a series of parliamentary sessions, each addressing a different political issue, the focus on a single slice of the full corpus further constraints the thematic range of the examined sentences, which, as outlined above, is beneficial to the task at hand.

With the 100k corpus forming the foundation of the data preparation pipeline, additional steps have been taken to ensure that the data ultimately used in training the final system is well-suited to the demands presented by the ID-reduction task. In the initial step, the domain of the source corpus was constrained even further, by removing outliers with respect to the majority of the corpus' content. This was accomplished by training the system-internal LM on the 100k corpus and subsequently calculating the model's perplexity for each individual sentence, followed by the removal of 10% of the corpus' sentences assigned the highest perplexity scores. The intuition behind this step is that a model trained on a domain-restricted corpus can be expected to exhibit worse performance on domain-atypical examples as opposed to ones representative of the training domain. For this step, 80% of the sentences comprised the training set, 15% were reserved for vali-

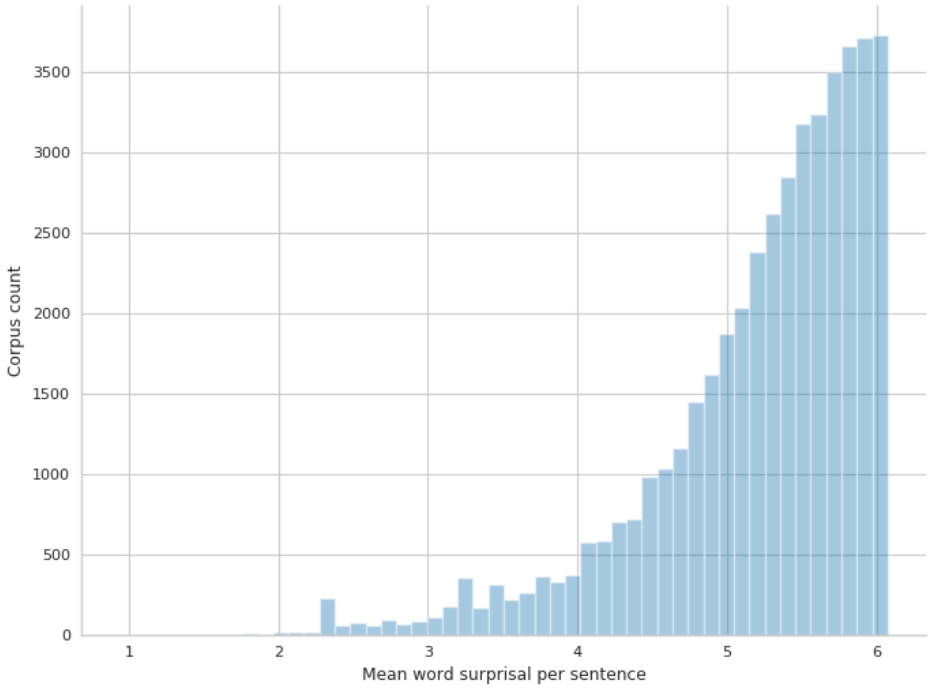
[3] This is done after filtering out sentences comprised of two or less words from the full Europarl corpus, so as to allow for more efficient mini-batching during training.

dation, and the remaining 5% formed the test set. So as to reduce the size of the LM's embedding matrix and expedite the training process, words containing numerals were first replaced with the <NUM> token, whereas words occurring within the training set with a frequency of less than 3 were substituted by the <UNK> token. The exact hyper-parameter choices for this initial pre-training of the IDGAN-LM can be found in Tab.1 supplied in the appendix.

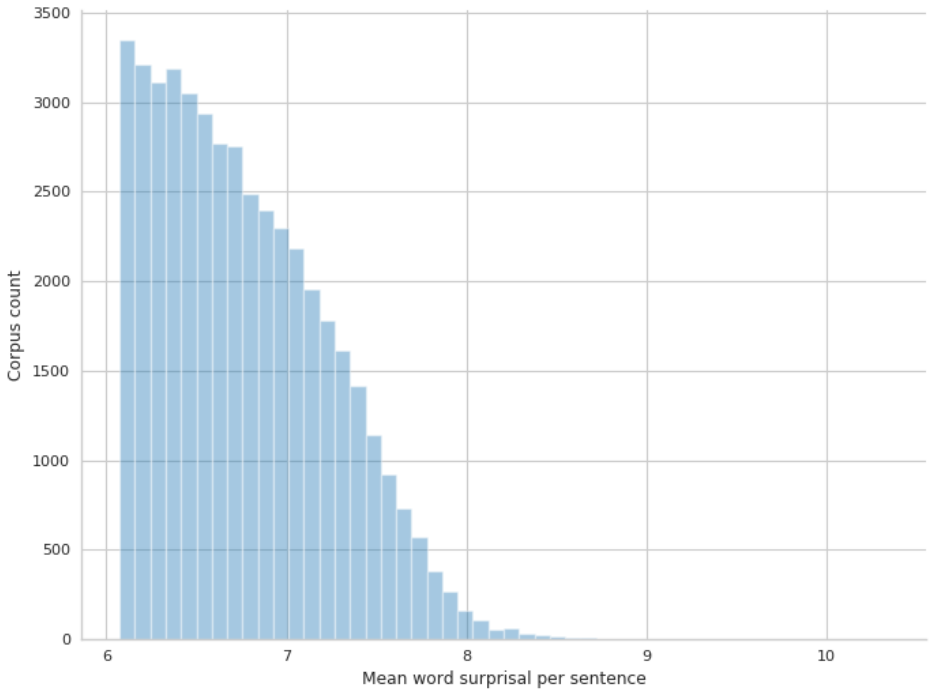
As the second step within the data preparation pipeline, the LM was re-initialized with random parameters and re-trained on the 90k corpus. Same modifications to the vocabulary were made as before, hyper-parameters remained unchanged, and data was split along the same proportions. Subsequently, each sentence within the 90k corpus was assigned the length-normalized surprisal score, calculated in accordance with equation (1). Following that, all sentences were ordered according to their surprisal value and the corpus subsequently split in two halves along the median score of 6.07. The two resulting sets contain 45k sentences each, with the corpus comprised of sentences below the median designated as 'low-ID' and the other assigned the 'high-ID' label.

In order to assess whether the so obtained ID-variant corpora adequately capture the desired qualities and are well-suited as training materials for the ID-reduction objective, a comprehensive analysis of the relevant corpus characteristics has been conducted. Here, an overview of the findings is given, accompanied by relevant data visualizations.

The evaluation of the surprisal scores within the two derived corpora yielded predictable results in light of the method by which the latter were constructed. At 5.21, the mean low-ID sentence surprisal is lower than the mean of the high-ID corpus, situated at 6.77. To obtain further insights into the differences between the distribution of surprisal scores across the two corpora, significance tests have been performed. As Fig.1 illustrates, the corpus-specific surprisal score distributions are highly skewed, which is consistent with the observation that scores were distributed roughly normally within the source 90k corpus, as can be seen in in Fig.2. For this reason, the significance of the inter-corpus difference in surprisal was estimated using the non-parametric Wilcoxon signed rank test. The two distributions have been found to diverge significantly with $p < 0.001$. In addition, the corresponding effect size was found to be large, with Cohen's $d = -2.5$ (i.e. < -0.8).



(a) Low-ID corpus.



(b) High-ID corpus.

Figure 1: Normalized surprisal scores in the induced ID-specific corpora.

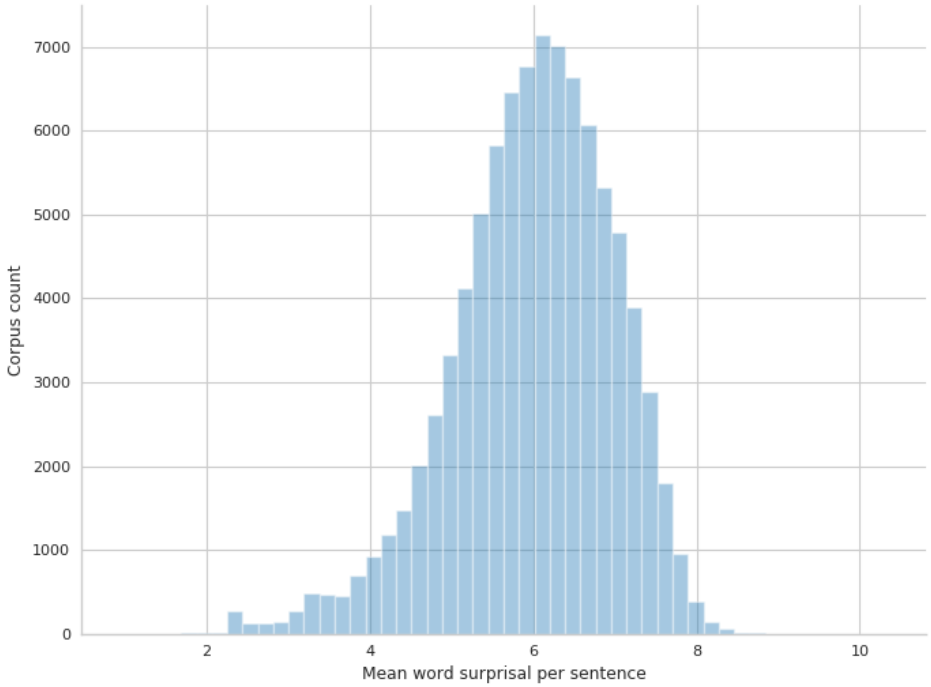


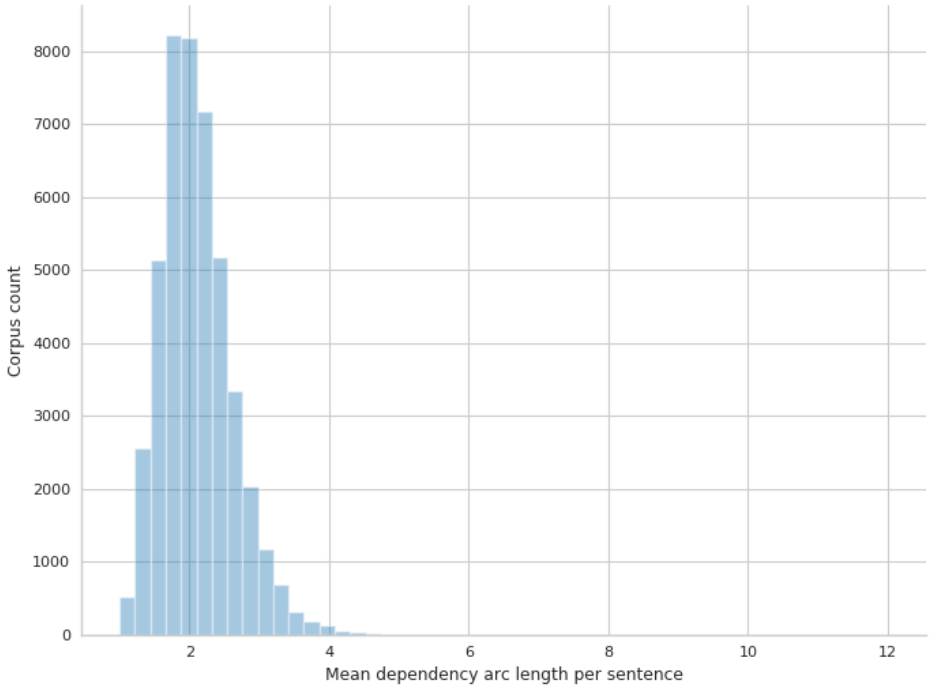
Figure 2: Normalized surprisal scores in the 90k-Europarl corpus.

From these observations it can be reasonably concluded that the ID-variant corpora do indeed effectively capture the surprisal differences between their respective constituent sentences.

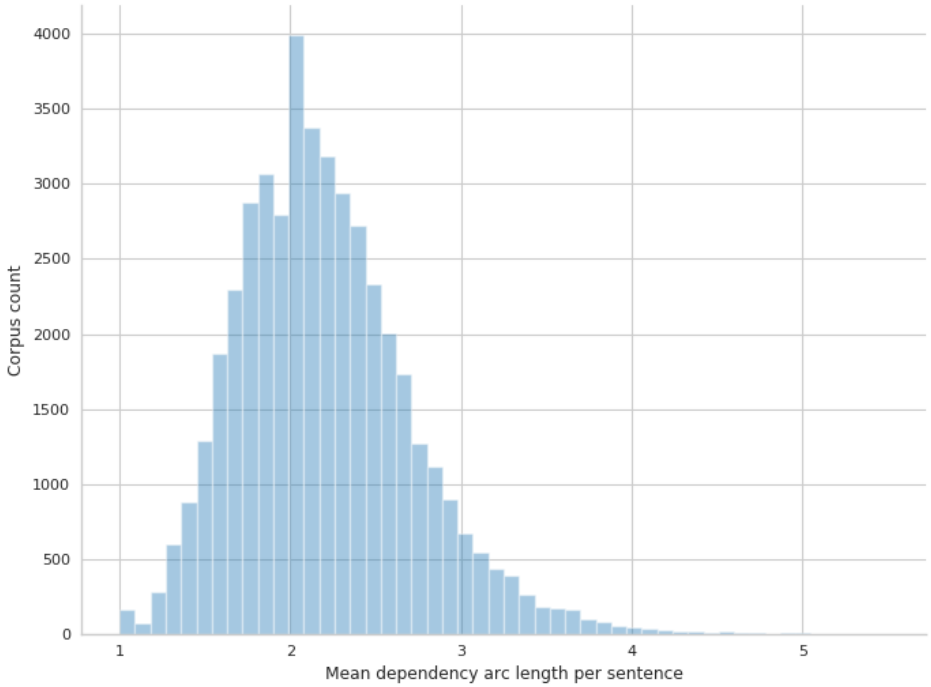
As surprisal, and more generally ID, is strongly correlated with processing difficulty in human comprehenders, it is of interest to examine whether other (psycho-)linguistic phenomena which have previously been shown to have a negative impact on processing speed are also captured within the ID-variant corpora. One such phenomenon is the length of dependency arcs describing the structure of a sentence, with shorter spans reducing the processing effort, as predicted by the Dependency Length Minimization hypothesis [15]. Another is the integration cost of discourse referents present within a sentence, as defined by the Dependency Locality Theory (DLT) [16]. Specifically, the theory posits that the processing effort associated with a sentence increases whenever new discourse referents are introduced between the head and the dependents of a dependency relationship, where 'discourse referent' corresponds to an entity or action that can be referred to with an anaphoric expression following its initial mention.

Both values can be estimated jointly with the help of an off-the-shelf dependency parser and POS tagger. In this work, the implementations provided by the SpaCy toolkit [28] are relied upon. While the cumulative dependency arc lengths can be trivially surveyed by measuring the distances between a parent and each of its children for all parent nodes in a sentence, the integration cost calculation requires an additional step. Since discourse referents are not marked as a result of dependency parsing alone, each sentence is first annotated with corresponding Penn Tree Bank POS tags, with individual nouns and non-auxilliary verbs designated as discourse referents in an approximation of DLT's assumptions. Subsequently, the combined integration cost at each parent node can be calculated as the sum of discourse referents encountered between the parent and its children, including each child node. The mean of the so obtained sum is then regarded as the approximate sentence-wise integration score, while the sentence-specific mean dependency arc length is established in a similar manner.

The results of the analysis conducted in this manner are as follows. Within the low-ID corpus, the mean sentence dependency arc length has been found to equal 2.1 words, which is lower than the mean dependency arc span within the high-ID corpus at 2.2 words. Similarly, the low-ID integration cost mean is lower at 2.27 energy units (with one unit corresponding to one intervening discourse referent) than the 2.4 established as the mean within the high-ID corpus. To see whether these differences are statistically significant and to estimate the associated effect size, an independent samples t-test was performed for each metric, as their distributions roughly fulfill the assumption of normality which can be seen in Fig.3-4 and have comparable variance. In both cases, the difference has been found to be statistically significant with $p < 0.001$ for dependency arc spans and $p < 0.001$ for sentence-wise integration costs. However, in both cases the effect size is small with Cohen's $d < 0.2$. As such, it can be concluded that the ID-variant corpora succeed at capturing meaningful differences with regard to the two examined criteria to a statistically significant, albeit minor, degree. More importantly, a positive relationship between variation in surprisal as manifested across the two corpora and other phenomena linked to processing difficulty can be established, which further validates the employed strategy. Given that a positive correlation between dependency length and surprisal scores had been established in the past [17], the observations detailed in this section serve as further evidence for the interconnected nature of differently motivated complexity metrics.

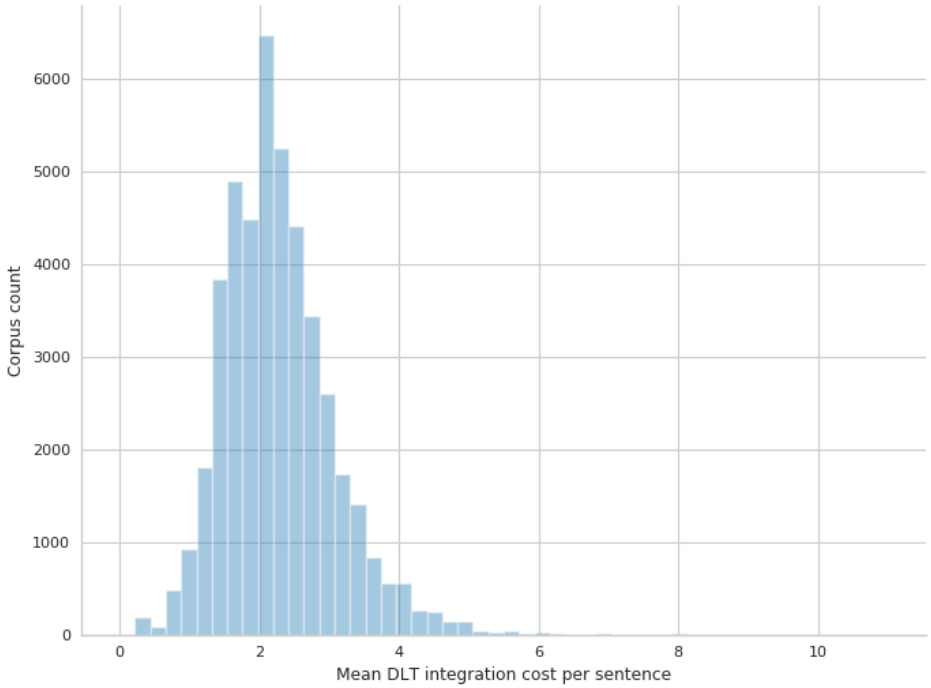


(a) Low-ID corpus.

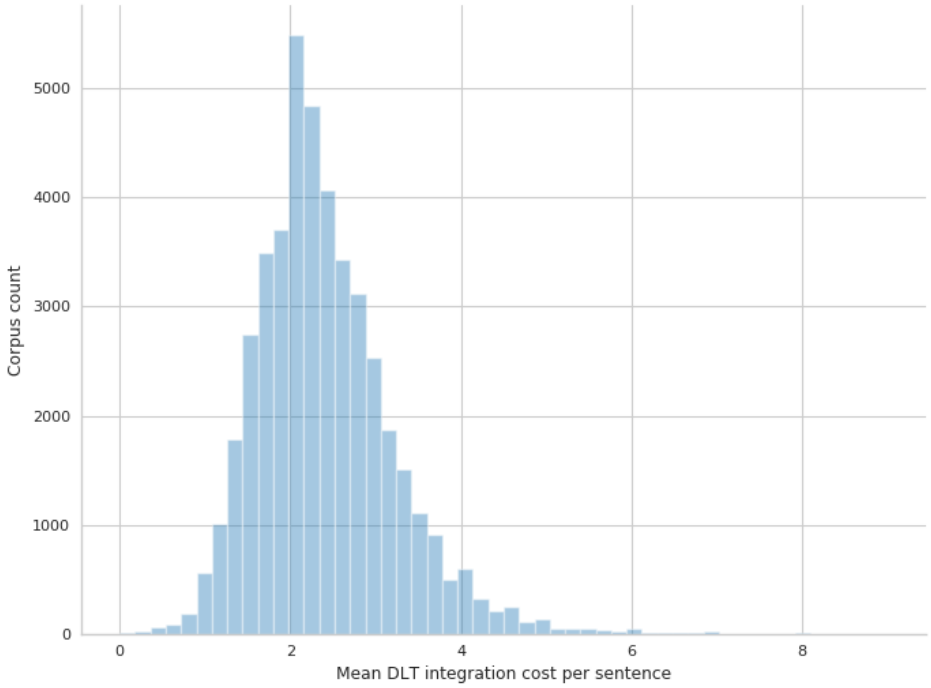


(b) High-ID corpus.

Figure 3: Mean dependency arc length in the induced ID-specific corpora.



(a) Low-ID corpus.



(b) High-ID corpus.

Figure 4: Mean DLT integration cost in the induced ID-specific corpora.

Lastly, attention was also given to the distribution of sentence lengths, POS tags, and corpus n-grams within both ID-specific corpora. Perhaps surprisingly, high-ID sentences were determined to be longer on average, with a mean length of 26.25 words, than sentences making up the low-ID corpus where the mean sentence length equals 24.36 words. This appears to go counter to the intuition that short sentences compress the communicated information to a greater degree which, in turn, should manifest itself in a correlation between reduced sentence length and increased mean surprisal. Moreover, the observed difference has been found to be significant ($p < 0.001$), but with a small effect size (Cohen's $d < 0.2$). While this inconsistency may ostensibly arise from the small scope of the examined corpora, it nonetheless presents an interesting finding, in that it lends further credence to the assumption that the ID-reduction task cannot be solved trivially by extending sentences to a greater length. Instead, a complex transformation targeting both the semantic and syntactic levels of a sentence presents a more likely solution. This further supports a fully data-driven approach to ID-reduction, as modeling a function which could reliably achieve this goal – especially on a large scope – via hand-crafted rules is both prohibitively challenging and time-consuming.

Differences in the POS tag distributions, on the other hand, are much harder to interpret. As Fig.5 illustrates, the divergences between the two corpora as regards individual tags are minute, without any striking patterns to be pointed out as part of a surface-level analysis. In either case, the relative proportions of the individual POS tags are roughly comparable, with only minor deviations. Summing over related tags, similarly does not yield any interesting insights as the distributions are, on the whole, very similar. The frequency distributions of 1-, 2-, and 3-grams within the low-ID and high-ID corpus also do not diverge significantly from another, as can be seen in Fig.6, instead closely following the trajectory predicted by Zipf's law in each case, as is characteristic for natural language texts. For none of the examined n-gram granularities can a consistent tendency towards assigning a disproportionately high number of low-frequency tokens to the high-ID corpus be observed.

The absence of notable differences for the latter two of the examined corpus characteristics is encouraging, as it suggests that the employed construction method for the ID-specific corpora, while naive, succeeds in isolating some corpus-internal characteristics which are

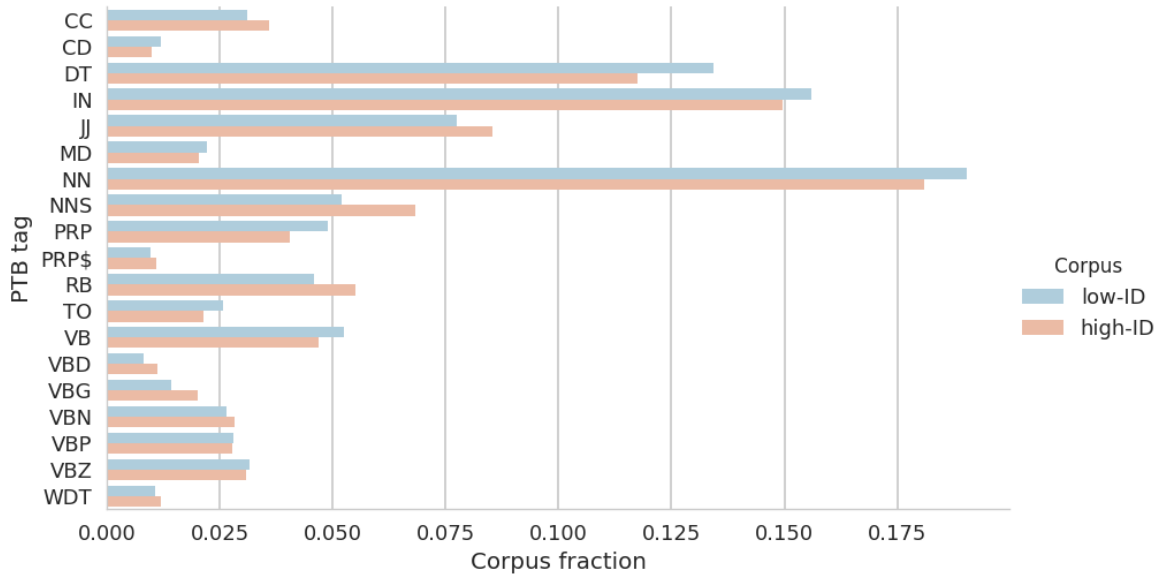


Figure 5: POS tags in the ID-specific corpora.

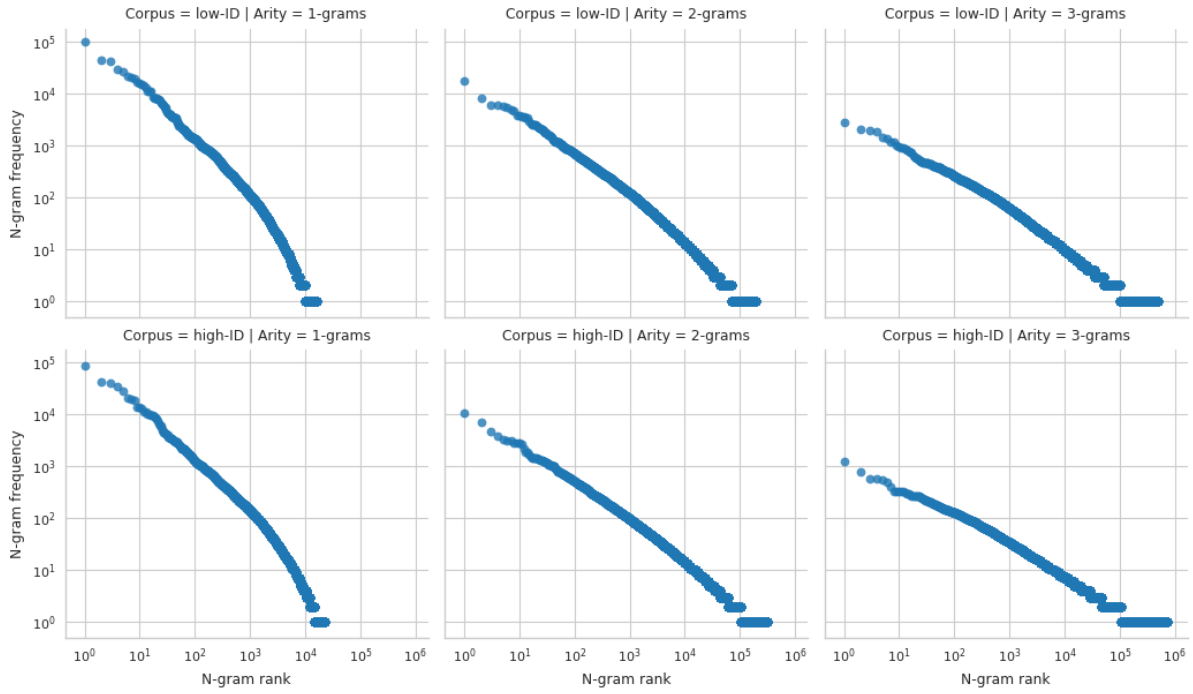


Figure 6: N-grams in the ID-specific corpora.

pertinent to processing difficulty – such as surprisal, dependency arc length length, and discourse referent integration cost – from those which are more general, such as POS tag and n-gram distributions. For that reason, the derived ID-variant corpora can be assumed to capture information which is required to provide an effective training signal for an ID reduction engine without encouraging linguistically anomalous or highly idiosyncratic output. Thus, they are deemed a good fit for the training of the proposed system.

3.2 System Design

The overall architecture of the IDGAN system, as touched upon in the earlier parts of this report, comprises several distinct, yet interdependent components, each implemented as a deep artificial neural network. This is motivated by the dual nature of the ID-reduction task, as its goals include the reduction of the surprisal score assigned to the input sequence in the generated output as well as the preservation of the content communicated by the input sequence. To adequately accommodate this multi-level objective, the ID-reduction engine relies on an SAE for constrained input reconstruction, an LM for word predictability estimates which serve as the basis for surprisal score estimates, as well as a classifier network to facilitate the GAN-assisted training. The latter is relied upon for establishing a mapping between the source distribution of high-ID sentence encodings and the target distribution of low-ID sentence encodings.

This section discusses the implementation of each component contributing to the IDGAN functionality, as the exact choices made in the architectural design of neural networks are one of the primary factors influencing their performance and fitness for the learned task. This overview is followed by a detailed description of the manner in which the individual components interact with each other at training time and during inference. Lastly, an analysis of the training process itself, including the optimization procedure, is conducted. The IDGAN system, as a whole, is implemented within the TensorFlow [1] framework, with the source code provided under *repos.lsv.uni-saarland.de/demelin/master-thesis*.

The IDGAN-internal LM, fulfilling the function of the surprisal estimator within the overall system, is a stacked RNN equipped with the LSTM cell, with each of its layers comprised of 512 individual cell units. The choice of the LSTM cell is motivated by its gated design which effectively discourages vanishing and exploding gradients, known to deteriorate the learning capacity of standard RNNs. It also has been shown to facilitate the learning of long-distance dependencies within natural language sentences by neural models. The bias of the 'forget' gate within the LSTM is initialized at 2.5 to further support dependency learning, in accordance with the training methodology employed in [33]. As speakers of English do not geneally have access to the right context in which a word occurs during incremental sentence processing, the LM is kept unidirectional,

only being able to access the previously seen sentence prefix. As alluded to before, the embedding table learned by the LM is simultaneously used to project the logits generated by the RNN during each prediction step into the vocabulary space. A softmax layer is then used to transform this projection into the predictive probability distribution. The tying of embedding parameters had originally been proposed in [46], where it was shown to improve the quality of the learned word embeddings and lower model perplexity.

All of the LM's parameters are initialized according to the initialization method introduced in [18], so as to further counteract the occurrence of extremely large and small gradients during the early training stages. To improve the LM's capacity for generalizing beyond the training data, dropout and L2 regularization is used during the training. The former randomly severs connections between the individual layers of the network, thus injecting noise into the propagated training signal and forcing the model to compensate for the incomplete information by learning to generalize. The latter, on the other hand, imposes as Gaussian prior on the distributions of the model's parameters, which has been shown to bolster the model's predictive power in addition to reducing the validation error.

Prior to the training of the full IDGAN system, during which its parameters remain fixed, the LM is trained using the ADAM optimizer [36] on a target-sampling objective defined in [32]. The choice of the objective is motivated by the large size of the target vocabulary, since candidate-sampling avoids the computational bottleneck associated with the calculation of a predictive distribution over the entire vocabulary, as is done when minimizing cross-entropy between model predictions and prediction targets without sampling. Throughout the training, the learning rate is decreased by 10% after each two consecutive epochs during which no improvement can be observed on the validation set. After 20 of such stagnant epochs, the training is terminated according to the early-stopping criterion.

The IDGAN setup incorporates a total of two SAEs, each used to construct informative, dense encodings of sentences drawn from the two corpora used in training and evaluating the system. For the sake of convenience and readability, the SAE tasked with encoding high-ID sentences and, as such, operating on the source domain, will be referred to as *translator SAE*. The desired outcome of the GAN supported training is for the translator SAE to learn the mapping function from high-ID inputs to low-ID outputs carrying the

same semantic content. The second SAE, tasked with encoding low-ID sentences, is referred to as *ground-truth SAE* from here on. The encodings it generates comprise the target distribution against which sentence encodings extracted from the translator SAE are evaluated by the discriminator network. Both SAEs are identical in their architecture due to the similar nature of their task, with the range of surprisal scores associated with their input being the only difference with regard to their respective operation. Owing to their intended functionality, both models consist of an encoder and a decoder network, with the final hidden state of the encoder used to initialize the decoder's hidden state. Motivated by the same reasoning as in the design of the LM, both component networks are implemented as two-layer, unidirectional RNNs equipped with LSTM cells and augmented with the global variant of the 'Luong' attention mechanism. In exploratory experiments conducted on small subset of the 90k Eurporal corpus a substantial positive impact of the attention mechanism's inclusion on the SAEs' input reconstruction performance could be observed.

Both SAEs are initialized with word embeddings learned by the LM during the corpus construction phase, which are subsequently fine-tuned during the IDGAN training, so as to better reflect the ID-variant nature of the data each SAE receives. Use of pre-trained embeddings is correlated with faster training speeds and had been found to generally improve model performance in the past. To improve the quality of the fine-tuned word embeddings, the embedding table of the encoder within each SAE is simultaneously used as the projection matrix for the outputs of the decoder RNN. Another point that bears mentioning is that sentences provided as inputs to each encoder are reversed, as this has been shown to produce better outputs in encoder-decoder systems [57]. As with the LM, the SAEs' parameters are initialized according to the 'Xavier' initialization strategy, while regularization is enforced through dropout and the L2 penalty. For an overview of the hyper-parameter settings used in training both SAEs as part of IDGAN, the reader is referred to Tab.2 in the appendix.

The last IDGAN component is the discriminator. While convolutional networks have become the default choice for discriminator architectures in adversarial training as applied to the visual domain, a fully-connected, feed-forward neural net is the natural choice for classification between two categories of sentence encodings represented by dense, one-

dimensional vectors. Within IDGAN, the discriminator is implemented as a three-layer network, where the initial layer has the same dimensionality as the sentence encodings generated by either SAE. The final layer, on the other hand, has a dimensionality of one and returns a scalar value for each input sample, which denotes its predicted class: low-ID or high-ID. To bolster the modeling prowess of the classifier, its design incorporates batch normalization [29] and layer normalization [2] which are used to the exclusion of the other, depending on the chosen adversarial objective. Both aim to normalize the neuron activations within each layer of the discriminator, but employ different strategies to achieve this goal. Batch normalization performs this operation over each mini-batch of input samples, whereas layer normalization does so over all summed inputs to a layer for each individual input sample.

Moreover, the network also incorporates parametrized rectified linear units [25] as the non-linearity of choice, since they, too, have been found to increase a neural net's expressiveness without adding a large number of extra parameters to be optimized. As before, parameter initialization within the discriminator is done in accordance with 'Xavier' initialization strategy, while dropout and L2 regularization are relied upon for the prevention of undesired over-fitting. Overall, the discriminator design is light-weight and simple, but, as the experiments performed as part of the present investigation will show, offers sufficient discriminatory power for adversarial training to be carried out. Additionally, the small parameter size is highly beneficial for a system combing multiple models, where memory limitations of GPUs are a consideration.

The complete IDGAN system which combines the individual constituent parts delineated above into a coherent whole, can be seen in Fig.7, while Fig.8 illustrates the system's intended functionality. Here, the discriminator forms the primary connecting point between the translator SAE and the ground-truth SAE. In this capacity, it receives as its inputs the sentence encodings generated by the SAEs' encoders on the basis of the ID-specific input sentences. The SAE's decoders, on the other hand, are not connected to the discriminator and are, as such, unaffected by the adversarial learning process. Within each SAE, the encoder and decoder networks remain connected, both via the passing of the hidden state from one to another as well as the decoder's access to the encoder's step-wise latent representations via the attention mechanism. Thus, both can be jointly

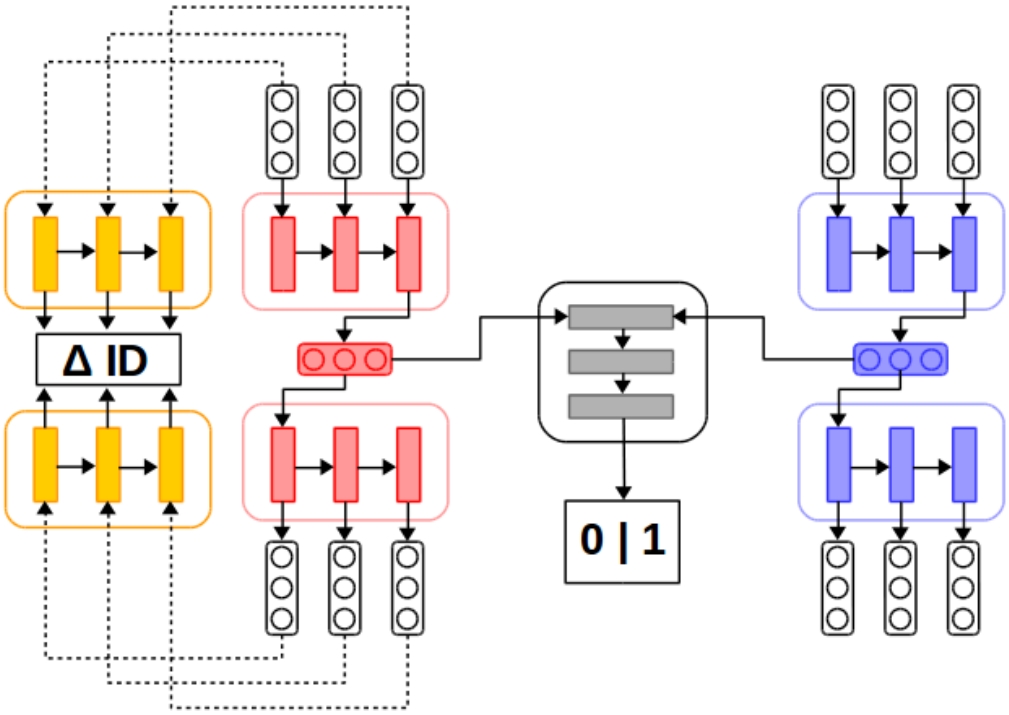


Figure 7: The proposed IDGAN system. Translator SAE is marked in red, whereas the ground-truth SAE is shaded blue. The central component represents the FFN classifier functioning as the discriminator. The LMs used to compute the obtained reduction in surprisal are colored yellow and used only during validation steps, as denoted by the dashed connections.

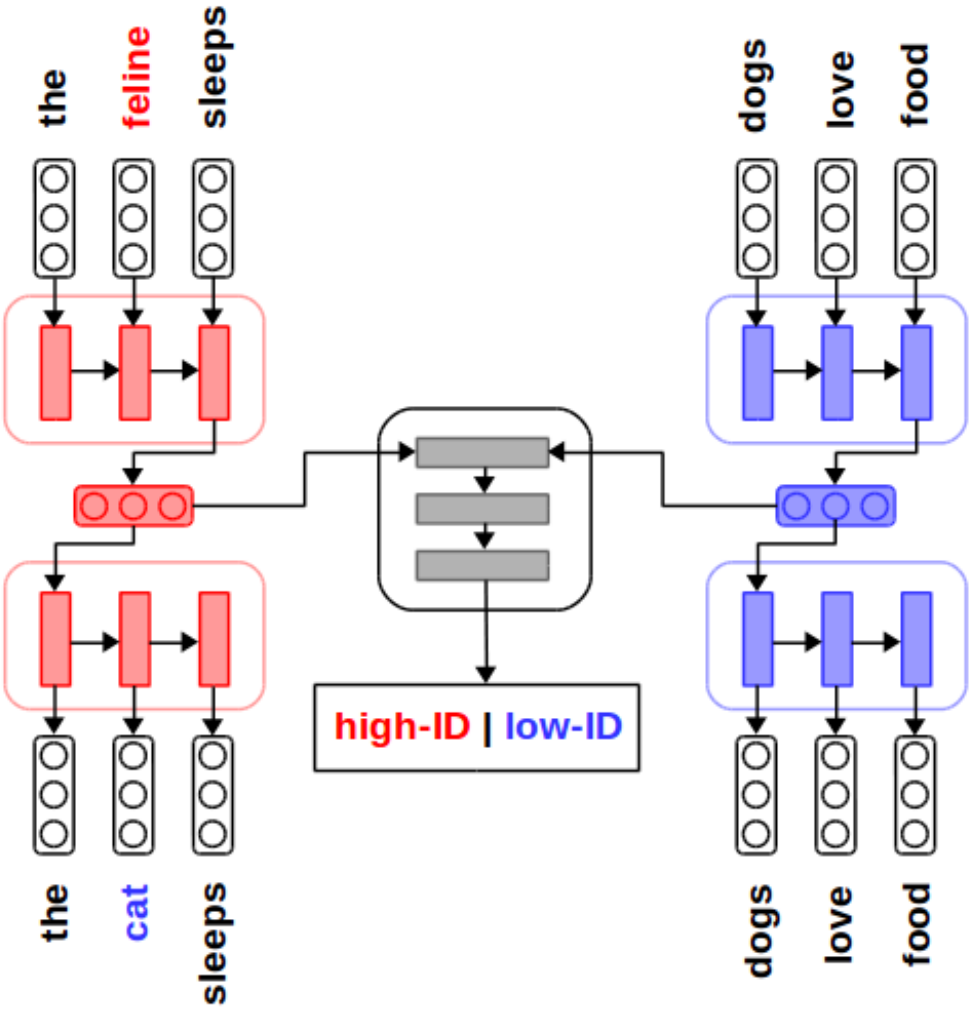


Figure 8: A simplified example of the proposed system's envisaged performance. High-ID sentences are fed to the translator SAE's encoder in order to obtain low-ID paraphrases from the corresponding decoder network. Importantly, the effect of applying the ID-reduction function to the input sequences is not limited to word substitutions, but also extends to complex syntactic and semantic transformations.

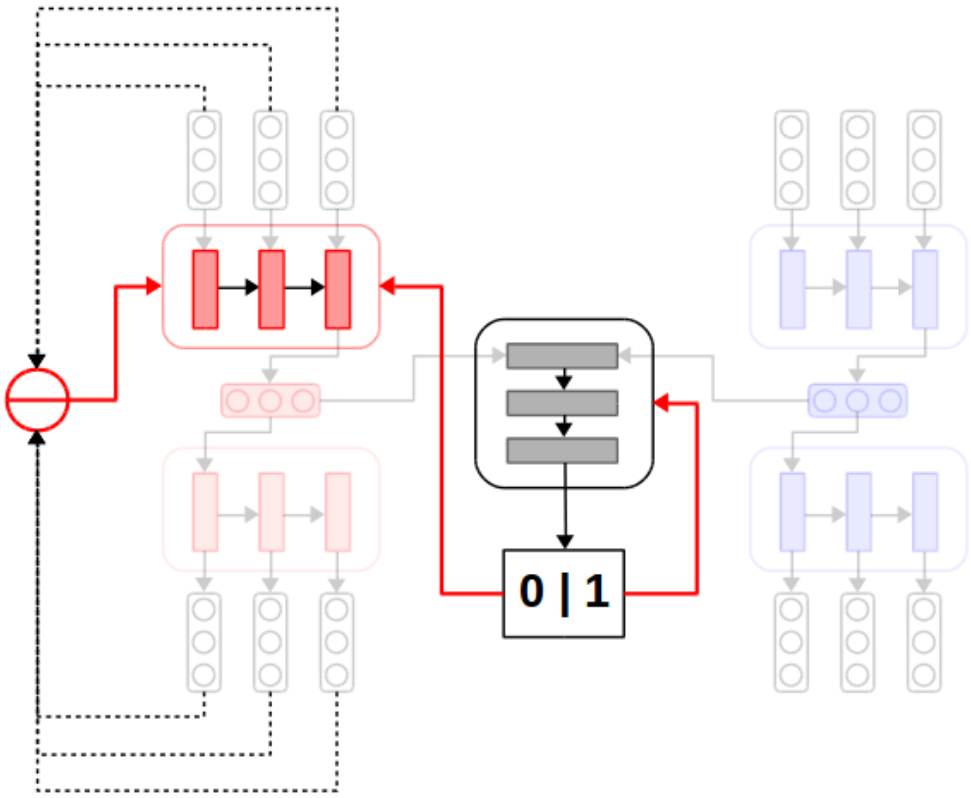


Figure 9: Error signal propagation within the IDGAN system. Red arrows represent the gradient flow used to train the highlighted components. Both the generator and the discriminator are trained on the discriminator's classification error, while the generator additionally learns to produce encodings from which its input can be reconstructed.

optimized via reconstruction loss as applied to the output of each decoder network, should this be desired. Moreover, the reconstruction loss can also be utilized to optimize only the encoder or only the decoder by constraining the set of trainable parameters within the IDGAN implementation prior to the commencement of training. The gradient signal flow used in the training of the evaluated IDGAN configuration is depicted in Fig.9.

The role of the LM, meanwhile, is to estimate the reduction in ID achieved by the system on the sentential level, at each training and evaluation step, by means of the surprisal metric. To do this, two LM instances are created and initialized with values learned on the 90k Europarl corpus prior to the assemblage of the ID variant corpora, as described in section 3.1. Doing so guarantees that the models had been exposed to sentences from the high-ID as well as the low-ID spectrum, therefore allowing them to make informed, domain specific decisions when calculating word predictability scores. In both cases the LM's parameters remain unaffected by the IDGAN training. At each step, one LM reads in the translator SAE's input sequence, while the other is fed the corresponding output sequence. On the basis of the word probabilities assigned by the each LM to the individual word tokens within their respective inputs, a mean surprisal score is calculated for the SAE's input and output alike, in accordance with Eq.(1).

The difference between the surprisal scores assigned to the input and output sequences is regarded as the achieved ID reduction. In case of an observed increase in ID, the corresponding value is therefore negative. The LMs thus perform a purely evaluative function, with the goal of guiding the training process towards the successful learning of the ID reduction function relegated in its entirety to the adversarial training objective. The LMs, nonetheless, have an indirect impact on the outcome of the training process, as the calculated surprisal differential is used as a secondary early stopping criterion during the training, in addition to the training objective. Thus, the training loop terminates prematurely if no improvements with regards to both values are obtained on the validation set for a preset number of training epochs.

The training procedure takes place in two discrete stages, a pre-training stage and the adversarial learning stage. During the pre-training stage, the translator SAE learns to accurately reconstruct high-ID sentences. As a consequence, its encoder is expected to

learn to generate sentence encodings which adequately capture the semantics and syntax of input sentences. Since the so obtained sentence encodings are used to initialize the hidden state of the decoder, their informativeness has to be sufficiently high so as to enable an accurate reconstruction of the input, which is assisted by the decoder-side attention mechanism. The ground-truth encoder, on the other hand, is trained on sentences drawn from the low-ID corpus, following a similar intention.

On a technical level, the pre-training step is carried out through the use of the ADAM optimizer in conjunction with the same candidate-sampling objective as had been employed to train the LM as part of the corpus induction process. As has been the case for the LM, the learning rate is progressively annealed over the course of the training loop, which terminates once either the maximum specified number of training epochs has been reached, or a predetermined number of epochs failed to produce an improvement on the validation set with regard to the training objective. Once this goal has been achieved, the so learned parameters are frozen in case of the ground-truth SAE and remain constant for the remainder of the IDGAN training. Throughout the training, sentences are supplied to either SAE in binned mini-batches so as to expedite the training duration. To prevent either model from memorizing inter-sentential correlations, the contents of each mini-batch are shuffled.

During the adversarial learning stage, as the name suggests, the IDGAN's generator is trained on an objective which combines adversarial loss as defined in Eq.(2) with the reconstruction loss in a weighted sum. The corrsponding compound objective is given in Eq.(3).

$$J_{Compound} = \lambda_{rec} * RecLoss + \lambda_{adv} * AdvLoss \tag{3}$$

Here, λ_{rec} denotes the relative importance of the reconstruction loss for the overall training objective, while λ_{adv} does the same for the generator part of the adversarial loss. The discriminator, meanwhile, is trained only on the discriminator part of the adversarial objective. Both lambda values are among the hyper-parameters with the most observed

impact on the performance of the system, and their optimal values can be determined empirically.

At each training step throughout this phase, sentences drawn from from the high-ID corpus are once again fed into the translator SAE, while samples taken from the low-ID corpus are forwarded to the ground-truth SAE. As opposed to the pre-training phase, both SAEs receive their inputs simultaneously. After reading-in the input sequences, the SAE-internal encoders generate informative sentence encodings as they had been trained to do previously. These encodings are then extracted from the encoder's hidden states at the final time-step and forwarded to the discriminator, which receives batches comprised entirely of encodings produced by either the translator SAE or the ground-truth SAE.

The discriminator's task is to assign the correct labels to the batched contents, with 0 corresponding to low-ID and 1 denoting the high-ID sentence encodings class. The associated classification error – scaled by its weight hyper-parameter – provides the signal for the adversarial update of the generator network, i.e. the encoder of the translator-SAE, and the discriminator itself, in accordance with Eq.(2). These updates are performed asynchronously, as the convergence of the adversarial training may be influenced by the ratio between generator and discriminator updates, with the optimal update frequencies varying based on the nature of training data as well as the specific loss function used in adversarial training. While general guidelines for the selection of these hyper-parameters exist, in practice finding the right ratio continues to present a major hurdle for training GANs successfully [50]. The results obtained by the present study corroborate the extent of this challenge.

In conjunction with the adversarial loss, the system also calculates the reconstruction loss for the translator SAE, by passing the sentence representations generated by the SAE's encoder to its decoder and comparing the decoded word sequences against its input. The reconstruction loss – scaled by its assigned weight – can then be used to adjust the parameters of the entire SAE or each of its constituent networks in isolation. In the majority of the conducted experiments, only the encoder had been trained on the reconstruction loss, so as to guide and constrain the adversarial updates which also only affect the encoder in its role as the GAN generator. Updating the entirety of SAE's

parameters on the reconstruction objective has, in turn, been found to quickly overpower the adversarial signal, even when the reconstruction lambda is set to a low value. This is to be expected, as in this scenario the decoder learns to compensate for any changes done to the encoder's output via the adversarial updates.

A related strategy considered in some of the conducted experiments has been to initialize the decoder of the translator SAE with the pre-trained parameters extracted from the ground-truth SAE, as made possible by their identical structure. In doing so, the intention is to bias translator SAE's decoder toward generating low-ID sequences on the basis of sentence encodings received from the corresponding encoder, which the ground-truth SAE's decoder has been trained to do during the pre-training phase. To preserve this bias, the decoder is not optimized via reconstruction loss during the adversarial training phase. After considering various hyper-parameter settings, ultimately no positive results could be obtained for this training configuration.

The motivation behind the use of a compound training objective given in Eq.(3) is that, should the adversarial objective not be constrained by some enforced notion of similarity between the input and output sequences, the output cannot be expected to communicate the same message as was contained within the input, which follows from the non-aligned nature of the training examples. The choice of the reconstruction objective as the content preservation constraint, however, is inherently problematic, and likely to be one of the main contributing factors towards the inability of the IDGAN, as conceptualized in this work, to produce consistently positive results. Section 3.3 offers further insights into this direction for error analysis. Without the addition of the adversarial objective, on the other hand, the second training phase - i.e. the adversarial stage - is identical to the pre-training phase, which deprives the system from any possibility to learn the ID-reduction mapping. It should furthermore be noted that a training regime based around the multi-task learning paradigm [49], where the adversarial and reconstruction updates are performed asynchronously, could theoretically present a viable alternative to the compound objective utilized here. However, this training configuration did result in improved system performance in the exploratory experiments conducted as part of the present study.

One final item to be addressed at this juncture is an alternative adversarial loss objective considered in a series of experiments alongside the standard formulation given in Eq.(2) once the latter has been reliably established to result in unsuccessful training runs when employed in IDGAN's second training phase. The training objective in question aims to reduce the earth mover's (EM) or Wasserstein-1 distance between the source and target distributions. One specific variant of this objective which forces the gradients of the discriminator loss to be upper-bounded by 1.0, so as to satisfy the requirements for an accurate estimation of the evaluated distance, is known as 'Wasserstein GAN with Gradient Penalty' (WGAN-GP) [21] and has been applied to the natural language domain in the past with promising results [47]. Thus, it is of immediate interest to the present investigation. An in-depth discussion of the mathematical background of the WGAN-GP objective is given in the cited publication, to which the interested reader is referred. At this point it suffices to say that although the use of the EM objective in the present investigation resulted in markedly different training behavior of the proposed system, as compared to the standard GAN baseline, this alternative remained similarly unsuccessful in facilitating the learning of the desired translation function.

Concluding this section, it must be reiterated that the IDGAN architecture and training regime detailed in this section are reminiscent of the AAE design, in that both apply adversarial training to latent representations generated by an encoder embedded within an SAE. Despite their similarities, the two systems nonetheless differ in several aspects. First, the IDGAN target distribution is not represented by a data collection directly, but is constructed at training time by deriving dense sentence encodings via an auxiliary SAE from the target text corpus. The information-theoretical properties of said corpus' contents are relied upon to guide the adversarial learning process. Additionally, IDGAN architecture incorporates an LM component used to monitor the changes in surprisal effected by the translator SAE. Importantly, the current system is applied to the natural language domain and has therefore to surmount the challenge of training GANs on discrete data, which, in itself, is a daunting task that holds great potential for future advances in NLP once it is better understood. The problem of ID-reduction, too, is markedly different from the ones addressed as part of the AAE investigation. In light of these system-level and problem-specific contributions, the IDGAN system can be regarded as the product of an independent investigation into the unsupervised modeling of automated ID reduction.

In this, it is jointly motivated by insights from research fields such as psycholinguistics, linguistic theory, and natural language processing.

Ultimately, the proposed system, as formalized here, has been found incapable of performing its envisaged function as an automated ID-reduction engine, as illustrated by the experiments discussed in the next section. The evaluation of the obtained, predominantly negative, experimental observations is followed by an analysis as to the factors potentially responsible for the system's overall failure to achieve its goals.

4 Experimental Evaluation

4.1 Pre-Training

In a multi-stage training setup, such as the one put forward for IDGAN, assuring the best-possible training outcomes at each intermediate training step is essential to obtaining a satisfactory performance for the overall system. In the present case, insufficiently pre-trained SAEs cannot be expected to produce sentence encodings which succeed in capturing the semantic and syntactic content of source sentences to a degree required for the adversarial learning of the ID reduction function. Similarly, should the IDGAN-internal LM exhibit high perplexity on the 90k Europarl corpus, it is unlikely to perform adequately when tasked with predicting word predictability scores for ID-variant corpus construction or for the purpose of evaluating the extent of ID attenuation accomplished by the translator SAE. Moreover, as both SAEs are initialized with word embeddings learned by the LM on the 90k Europarl corpus, it is important to assure that the learned embeddings are of a sufficiently good quality to serve as the foundation for the generation of ID-specific sentence encodings.

For this reason, a series of experiments was conducted following each individual training stage, including the induction of the high-ID and low-ID corpus, so as to eliminate the inadequacy of pre-trained IDGAN components as a potential source for any observed deficits in the system's ultimate performance. Here, an overview of this incremental evaluation process is given. As detailed in section 3.1, during the construction of the corpora used in conjunction with the SAEs, the LM was trained over the course of two separate iterations – first on the 100k Europarl corpus and, after pruning outliers perplexity-wise, on the related 90k corpus. In the former case, the model achieved a perplexity of 75.65, whereas for the latter case a validation set perplexity of 68.9 and a test set perplexity of 73.4 can be reported. To the author's knowledge, no published language modeling study has used the exact subset of the monolingual English Europarl corpus as training material, so that no direct comparisons to other models reported in extent literature can be offered. One possible, albeit not ideal, point of comparison are LMs trained on the Penn Tree Bank dataset, which is used frequently for estimating LM performance and, at 4.5 million

words, is roughly twice the size of the 100k Europarl corpus. Several of such models are evaluated in [46] and resemble the IDGAN-internal LM in their use of tied embeddings. However, the combined discrepancy in corpus and model parameter size does not permit a truly meaningful comparison, other than the observation that the reported perplexity scores are generally within the same order of magnitude as ones achieved reported here.

Additionally, the quality of the embeddings learned by the LM on the basis of the 90k corpus was estimated using the word pair relationship methodology outlined in [42]. Manual adjustment of the questions to the corpus resulted in thirteen relationship types with five to ten word-pairs defined for each. The exact list is provided as part of the IDGAN implementation. The overall accuracy of the predictions made on the basis of the extracted embeddings is low at 6%, which is roughly consistent with findings reported for RNN-LMs in [42]. A manual inspection has furthermore revealed that in the vast majority of cases the predictions are synonymous with one of the entities defining the relationship, one example being 'Germany is similar to Berlin as [Germany] is similar to Amsterdam', where the predicted word is in denoted by square brackets. The predicted word is therefore semantically related to the target, 'Netherlands', as both represent geopolitical locations. This, in conjunction with the ubiquity of predictions following this pattern, suggests among other things that semantically unrelated words do not tend to be closely clustered within the embedding space, as this would have likely made the predictions mode semantically divergent. The insights thus collected from the two-pronged evaluation of the LM, while not unambiguously attesting to its adequacy, at the very least suggest that the observed performance does not significantly deviate from the expected performance of comparable neural models of language.

Following the completion of the pre-training stage, both SAE's have also been evaluated as stand-alone models isolated form the IDGAN system, so as to ascertain their general capacity for learning informative embeddings, as applied to the task of sequence reconstruction. During the pre-training both models had been exposed to equal quantities of data. However, for its training to converge to a low error rate, the translator SAE's batch size had to be halved, as compared to that used to train the ground-truth SAE. As such, the training procedure took twice as long in the case of the former. This can be taken as evidence for the greater difficulty posed by the corresponding optimization problem.

To obtain a meaningful estimate of the reconstruction accuracy for each SAE, the mean BLEU score was calculated on the test set corresponding to the ID domain assigned to the examined SAE, e.g. low-ID for the ground-truth SAE. The BLEU estimate was obtained using the *multi-bleu.pl* script distributed with the Moses machine translation system[4], as is common for NMT studies[6]. The script calculates the mean n-gram (for 1-, 2-, 3,- and 4-grams) overlap between the original test set and its reconstruction generated by an SAE, penalizing the relative importance of shorter overlapping spans inf favor of longer text stretches. The translator SEA achieved a BLEU score of 72.24 on the high-ID test set, while the ground-truth SAE was assigned a score of 78.93 on the low-ID test set. Given that a perfect reconstruction would correspond to a score of 100, the achieved reconstruction quality can be reasonably deemed high and adequate for the learned task.

A manual inspection of the reconstructed sequences supports this conclusion, with shorter sentences being reproduced perfectly in most cases, whereas longer sequences tend to diverge more from their targets, which is consistent with previous findings [57]. Interestingly, applying the pre-trained translator SAE to the low-ID test set produces a markedly high BLEU score at 86.68, as compared to the reconstruction quality displayed by the ground-truth SAE on the same test set. This is to be expected given the less predictable quality of sentences comprising the high-ID training corpus, as reflected by their higher overall surprisal, which may have required the translator SAE to learn more complex patters, leading to an improved capacity for generalization. As can be expected, running the ground-truth SAE on the high-ID test set results in a comparatively low BLEU score of 62.61. Complementing the previous observation, this indicates that a lower magnitude of modeling power is sufficient to accurately reconstruct the more predictable sentences populating the low-ID corpus. The discrepancy between the capabilities of both pre-trained SAEs observable in the cross-evaluation setting suggests that training encoder-decoder models on high-ID sentences can be expected to result in more robust models with greater capacity for generalization beyond their training domain, at the cost of longer training times. For the purposes of the IDGAN training, however, the performance of both SAEs has been deemed satisfactory, due to the near identical BLEU ratings obtained on their assigned ID-variant corpora.

[4]github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

4.2 IDGAN Experiments

The training of IDGAN did not result in the desired outcome and, indeed, was unsuccessful at accomplishing any degree of reduction in surprisal between the generated output and the received high-ID input sentences. As such, a traditional quantitative and qualitative evaluation of system's performance is not an option that is open to the present study. Thus, rather than evaluating the statistical significance of the observed ID reduction effect or the content preservation during the translation process, the evaluation of the IDGAN system follows a different route. As part of the investigation into the general capacity of the IDGAN to learn the ID reduction mapping, a series of exploratory experiments was performed. On average, each was carried out for roughly 30 epochs and was meant to assess the impact of different hyper-parameter configurations of the system's behavior. The chosen epoch window was deemed to be sufficiently indicative of the training's trajectory following the initial experiments. The so manipulated hyper-parameters included 1. the number of generator updates during each training step, 2. the number of discriminator updates during each training step, 3. the weight assigned to the adversarial loss within the compound generator objective as defined in eq. (3), 4. the weight of the reconstruction loss within the same objective, and 5. the adversarial objective itself. In the following a brief overview of the collected observations is given and conclusions are drawn which may benefit future applications of GANs to NLP problems and the task of ID reduction, specifically.

Even though, as mentioned in the previous sections, three adversarial objectives are considered in total, the focus of the conducted experiments lies on the likelihood-centered objective outlined in eq. (2). Fig. 8 provides a visual summary of several of the most interesting hyper-parameters configurations iterated over during the experimentation process, illustrating their effects on generator loss, discriminator loss, and the achieved ID reduction. Representative output samples generated by the translator SAE over the course of the entire training period are supplied together with the IDGAN source code, but are near-uniform in their unnatural quality. As the adversarial losses and their respective relationship to each other are difficult to interpret, especially when applied to a novel task in a domain foreign to the original conceptualization of the employed method, the examination undertaken here is strictly empirical in nature and focuses on the interaction

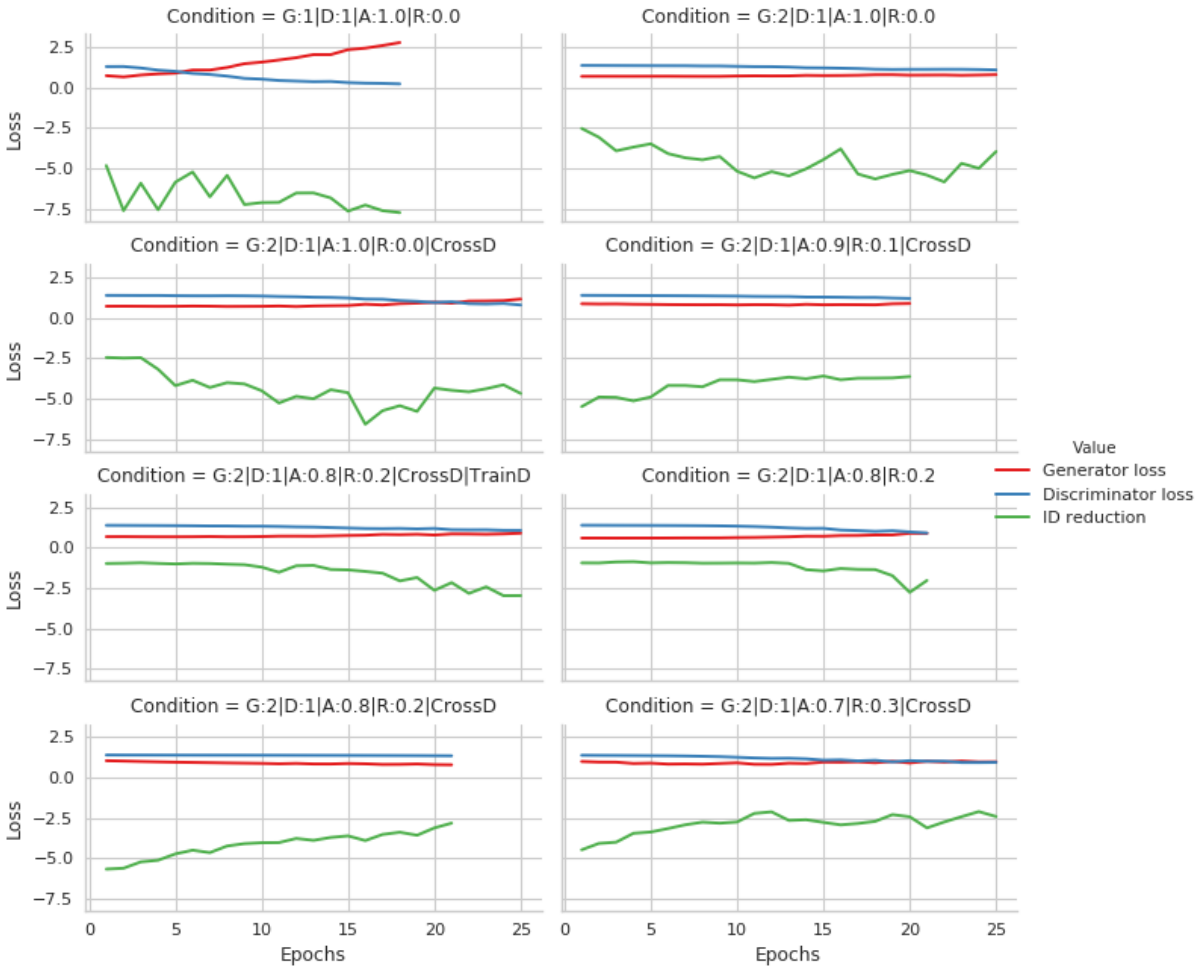


Figure 10: Validation set loss and ID reduction trajectories for IDGAN training with the likelihood objective.

between the loss trajectories, the chosen hyper-parameter settings, and the corresponding reduction in surprisal scores.

Within fig. 8, each facet of the plot depicts one experiment, with the corresponding hyper-parameter configuration denoted by its title, with D standing for discriminator updates, G denoting generator updates, A standing for the weight of the adversarial loss within the generator objective, and R standing for the reconstruction loss weight within the same compound objective. For each experiment, the validation scores corresponding to the initial 30 training epochs are shown, except in cases where training was interrupted prematurely due either technical malfunctions or constraints on time and the available computational resources. No validation was performed during the 'warm-up' period comprising the initial 5 training epochs. All experiments were performed on a single Nvidia TITAN X GPU and ranged in duration from 6 to 10 hours.

As is readily apparent from the presented objective trajectories, the translator SAE does not succeed in learning to reduce the ID of input sentences over the course of the examined training period. Nor is it possible to detect a clear tendency towards such an outcome occurring over longer training durations, as the ID reduction curve never goes over into the positive and tends to stagnate after varying numbers of initial epochs. Nonetheless, the so collected data does offer a number of interesting results which, should they be followed up upon, may contribute to a working version of the proposed system. Since a full grid-search over the five-dimensional hyper-parameter space would be prohibitive in regards to both time and computational resources required, the experiments under consideration were conducted following an 'informed' grid-search strategy.

First, an exploratory query into the update-specific relationship between the generator and discriminator within the IDGAN was launched, by training the system on adversarial loss alone and a balanced update schedule, i.e. each single generator update was followed by a single discriminator update. The so obtained observations can be seen in the top-left facet of fig. 8. From the loss trajectories, it is clear that the discriminator quickly 'overpowers' the generator, with the adversarial loss of the former going towards zero after ˜5 epochs and that of the latter increasing in parallel, with the divergence becoming more pronounced as time goes on. To address this pathology, the discriminator update

frequency was doubled, resulting in the losses converging steadily, as is ideal for the defined objective, with the most illustrative example found in the lower left corner of the figure. As a consequence, the 2:1 update ration was retained for the remainder of experiments. Moreover, these initial experiments demonstrated clearly that some degree of enforcement of content consistency between IDGAN's inputs and outputs appears to be required, as even when the GAN training converges, ID reduction remains low. As a manual inspection of the generated samples from the associated conditions has shown, this is a consequence of the translator SAE outputting sequences of single word token repetitions, e.g. 'the the the ... ', similar to the output of an SAE in the early trainig stages.

With the adversarial loss component thus stabilized, the focus was subsequently shifted towards addressing the the evolution of the reported ID reduction performance. To do so, three previously mentioned conditions were queried, where the decoder attached to the generator was either initialized with it's pre-trained parameters (default), the parameters of the ground-truth decoder with the intention to impose a bias towards the decoding ow low-ID sentences (CrossD), or where the cross-initialization was followed by further raining of the decoder on the reconstruction objective (TrainD). Neither of these strategies has ultimately born fruit. In the default configuration, when the reconstruction lambda is set to be low, at 0.2, the reconstruction loss is at first dominant, assuring a large degree of overlap between the translator SAE's inputs and outputs, with the corruptions caused by the initial stages of the adversarial training leading to a relative increase in ID within the generated output. Once the adversarial signal is string enough, i.e. upon the convergence of the generator and discriminator losses, the reconstruction quality quickly deteriorates due to a shift in the encoder's output distribution, yet the adversarial transformation does not bring with it the desired ID reduction effect, as the facet in the third row of the second column shows.

This trend is reversed when the decoder is initialized with parameters from the SAE trained to generate low-ID sentence encodings. Initially, the reconstruction quality is low, which is unsurprising, given the low BLEU score reported in section 4.1 for the ground-truth SAE when applied to the high-ID test set. As the training progresses, a downward trend with respect to the ID increase can be observed. The lowest row of the

plot provides insight as to the factors responsible for this development - the trend is more pronounced in the condition where the reconstruction loss is given more weight, thus it can be concluded that the ID increase is reduced as the input and output sequences become more similar. Importantly, at no point is the relative difference in ID positive, meaning that any potentially enforced low-ID decoding bias is of little effect to IDGAN's overall performance.

When the crossing operation is followed by decoder training on the reconstruction objective, the observed patterns resemble those found to hold for the default condition, which implies that the reconstruction signal is much stronger than the adversarial signal in the initial training stages, especially. One possible reason for this is that the translator-SAE is pre-trained on the reconstruction objective and thus already optimized for this partial task, with the pre-trained parameters corresponding to a local minimum within the reconstruction error landscape. As such a stronger adversarial signal is required for the generator to leave the minimum, which is immediately followed by a drop in ID as the generator's output quality starts to degrade. Overall, from the examination of IDGAN's behavior on the likelihood-centric adversarial objective, several conclusions can be drawn, including the positive correlation between reconstruction quality deterioration and increase in ID, the difficulty - perhaps impossibility - of balancing the adversarial objective with the reconstruction objective, and a potentially negative impact of pre-training the translator SAE on the reconstruction objective exclusively.

Once the adversarial loss formulation in eq.(2) was found to not yield any promising results within the IDGAN setup, a smaller set of experiments was carried out on the basis of the WGAN objective. A visual documentation of representative training outcomes is provided in fig. 9. One drawback of the WGANGP conceptualization is that the partial losses cannot be interpreted as straight-forwardly as in the previous case. This, in turn, makes stabilizing the adversarial training process more difficult, even when following best practices laid out by the research community. The latter has been done in the present study by reducing the learning rate for both GAN components to .00005 and setting the discriminator's - or, more accurately, the critic's - udates to occur substantially more frequent than those of the generator. Still, as fig. 9 illustrates, the observed component losses remain high, often exhibiting a large variance. At the same time, they

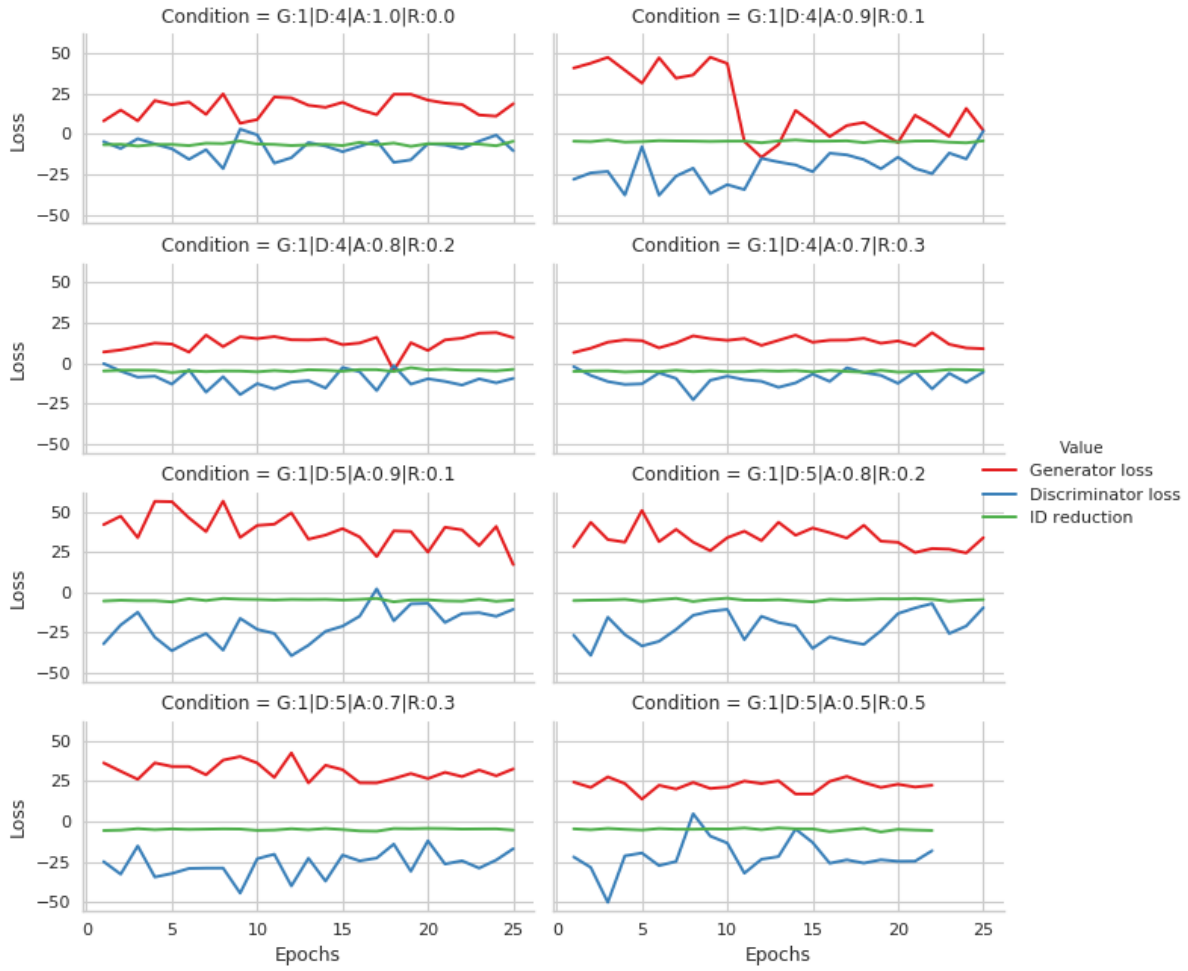


Figure 11: Validation set loss and ID reduction trajectories for IDGAN training with the WGAN objective.

appear to have little impact on the achieved ID reduction, which remains negative in all of the examined conditions and comparable in scale to the results obtained via eq.(2). As the initial experiments in the default configurations did not produce any interesting or intuitively interpretable outcomes, no crossed decoder initialization was performed. Despite the critic's loss approaching optimality, by being close to 0, in several of the tested experimental settings, which is expected to lead to an improvement in the generator's performance, the desired learning of an ID-reducing mapping did not take place. One possible explanation for the lack of success is the limited training duration, as WGANs are described in the original publication as requiring longer training periods to produce good results. However, when IDGAN instances equipped with this objective were let to train for longer stretches of time, the training loop was usually terminated after 50-60 epochs, following 20 validation epochs during which no improvements w.r.t. to the training objective or the extent of ID reduction could be observed.

4.3 Error Analysis

The possible reasons underlying IDGAN's failure are varied. Here, some of the ostensibly likely sources of the observed pathologies are put forward and briefly discussed. It should, however, be noted that without further experimental insights such suggestions remain primarily speculative in nature.

Although the method by which the ID-variant corpora had been constructed, as described in 3.1, was intended to steer the adversarial learning process in the correct direction, it may have been either ill-suited or not sufficiently effective for the tackled problem. One central weakness of the employed approach is that the bulk of sentences are only minimally different with regard to their surprisal ratings across the two derived corpora. This is most apparent when examining figures 1 and 2. which depict the split of the source corpus - the 90k truncated English Europarl subset - along the median of the corresponding surprisal distribution. As the surprisal scores are distributed normally in the source corpus, most of the sentences within each ID-variant corpus lie close to said median value. Conversely, it is the outliers occupying the tails of the low-ID and high-ID distributions which contribute most to the, as previously established, statistically significant difference in surprisal scores

between the two derived corpora. This is problematic in several ways. On the one hand, the for the most part minimally different samples making up the bulk of each corpus are unlikely to provide a sufficiently strong signal as as to guide the adversarial training towards the direction of ID reduction, due to its overall low magnitude. As such, even by attempting to control for other sources of linguistic variation such as the domain of the used texts, isolating surprisal differences above other interfering factors remains difficult.

Furthermore, the disproportional contribution of outliers towards the observed divergence between the two surprisal distributions is undesireable, as such sentences are not representative of the majority of sentences originating from the same source. This, in turn, limits the usefulness of the training data further, with information that is most valuable for the training objective being sparse and thus difficult for the translator SAE to generalize from, made even more difficult by the inherently sensitive and unstable nature of adversarial training. Moreover, it should be noted, that controlling for highly lexical characteristics such as register, domain, and vocabulary choice does little to isolate syntactic surprisal, which would necessitate the inclusion of syntactic information into the corpus construction process. In what exact way this could be achieved poses by itself an interesting research problem. All in all, the chosen approach to corpus construction does not appear to contrast differences in surprisal between the obtained corpora to a sufficiently large degree which would enable it to successfully guide the adversarial training procedure.

Another challenge the IDGAN proposal struggles with is the high number of hyperparameters, the optimal values of which need to be empirically established so as to guarantee the best-possible performance. This set includes the number of generator updates and discriminator updates, adversarial and reconstruction loss, the definition of the adversarial objective, but also task-agnostic factors such as the learning rate, choice of the optimized, dimensionality of each component network, regularization strength, and the choice of exact network architectures. For instance, a recursive neural net might achieve better results on the ID reduction task, as hierarchically derived sentence representations can be expected better capture syntactic variation indicative of surprisal effects. Moreover, while the interactions between some of these parameters are more or less clear, as is the case for the generator and discriminator updates, other are less obvious and difficult

to detect and, consequently, control for. Barring an extensive exploration of parameter choices in either brute-force or a more informed manner as done here, a promising approach to ameliorating this complexity issue would be to perform hyper-parameter search automatically [56] - at great computational expense - or, otherwise, drastically cut down on the IDGAN's system's complexity.

On the other hand, chief among the limitations restricting the capacity of the IDGAN to generate output which is both well-formed and sufficiently varied so as to model phenomena commonly associated with low surprisal, is its reliance upon input sequence reconstruction as the auxiliary objective used to stabilize the adversarial learning procedure, by restricting its hypothesis space. While this does indeed enforce a surface-level consistency between the input and output of the translator-generator SAE, this is accomplished on a token-by-token basis, minimizing the cross-entropy between the received target token and the generated output token at each time step. This, however, makes it difficult for the system to learn any useful syntactic transformations as part of the ID reduction process, whenever the contribution of the reconstruction loss to the overall training objective is above some threshold, leading to its dominance over the adversarial loss. This, naturally, is deeply problematic, as the more canonical syntactic structures are associated with a lower surprisal scores than the less frequently encountered constructions. For the ID reduction engine to be capable of translating high-surprisal sentence-level structures into their corresponding, low-surprisal counterparts – while also preserving the semantic meaning of the translated sequence to a large degree – is essential for achieving good performance on the ID reduction task. The dependence on reconstruction loss, however, effectively prohibits such transformations to be learned from data alone. The reliance upon sentence embeddings to enable adversarial learning, too, may be problematic, since, as [61] show, the decoding of well-formed sentences from arbitrary sentence encodings - such as are obtained via the transformation of high-ID encodings by IDGAN's generator - is a problem that requires tailor-made solutions and cannot be trivially solved by added reconstruction objectives or biases imposed on the decoders parameters.

This concludes the analysis of IDGAN's performance on the task of ID reduction as documented in a series of exploratory experiments. While no positive results could be obtained, the collected observations nonetheless offer some interesting insights into applying

adversarial training to an NLP task motivated by information-theoretical considerations, its limitations, and failure cases. Furthermore, suggestions were made in how some of the encountered pathologies may be addressed effectively, as part of future research. An immediately relevant extension of the IDGAN system presently in the works is detailed in the next, pen-ultimate section.

5 Future Research and Paths to be Explored

One of the envisaged goals at the outset of this work has been the development and evaluation of a general-purpose ID-reduction engine. While the results have ultimately fallen short of this ambitious goal, the collected observations have nonetheless offered a number of valuable and interesting insights. Moreover, the proposed system, in spite of its shortcomings, can readily serve as a useful point of departure for future unsupervised systems seeking to address the research problem of automated ID reduction. The negative results obtained following the completion of the extensive set of experiments described in the previous chapters should provide some guidance as to which directions are likely to prove fruitful and which hold less promise when it comes to system architecture design and the parameter choice, in particular. With this in mind, one potentially promising direction for future research lies in addressing the use of reconstruction loss as part of the primary IDGAN training objective through modifications to the chosen training regime and system design. In the following, a brief survey of steps towards this goal, which are being actively explored at the time of writing, is presented.

As part of an investigation into how this shortcoming may best be addressed, an alternative auxiliary task has been considered to replace reconstruction as the content-preserving training criterion. From the above problem description, an obvious solution is to enforce content similarity on the sequence-level rather than per word token, as that should allow for syntactic permutations to be learned more effectively. This naturally follows from the variability of natural language expression, as there usually exist multiple ways to express the same meaning proposition, with each such way exhibiting a different syntactic structure with a distinct syntactic surprisal rating. To incorporate an auxiliary objective into the proposed system which satisfies this requirement, a sentence similarity classifier has been implemented, based on the work presented in [43]. Within an extended ID reduction engine, such classifier can be used to estimate the similarity between the input and output sequences of the translator SAE. Minimizing their dissimilarity in parallel to optimizing the system's parameters on the adversarial learning objective, e.g. by combining both within an importance-weighted sum, should provide the augmented system with a means to preserve the semantic content of any given input sequence. By observing this constraint

during training, the system can be expected to learn permissible syntactic transformations leading to a reduction in mean sentence surprisal, assuming that such a strategy would contribute towards the reduction of the generator portion of the adversarial training loss.

Within the sentence similarity classifier, the similarity score is calculated as the Manhattan distance between the meaning representations of its two input sequences, i.e. the input and output of the translator SAE. These compressed sentence encodings, in turn, are obtained from two encoder LSTM-RNNs sharing their respective parameters. The choice of this specific model as a candidate for the modeling of the consistency-enforcing auxiliary objective is motivated by its near state-of-the-art performance for sentence classification combined with its comparatively small size, as satisfactory classification results can be obtained with a hidden layer dimensionality of 50. To ascertain that the classifier is compatible with the proposed system, several preparatory steps are required. In accordance with the strategy employed in the original publication, the classifier is first pre-trained on a human-annotated similarity corpus. For this purpose, a combined sentence similarity corpus is used, which encompasses the SICK corpus as well as data used in the SemEval 2013 Semantic Textual Similarity task[5].

Prior to combining these two resources, the SICK corpus is extended with synthetic examples which are generated via the WordNet-based synonym substitution method outlined in [7], which effectively doubles it size. While the so obtained extensions are quite noisy, this strategy, when employed, has been nonetheless shown to be effective for the training of sentence classification models. Furthermore, the synthetic data increases the robustness of the learned discriminative model, which greatly benefits the subsequent transfer learning step required to adopt the classifier for the Europarl domain. The implementation of the classifier provided together with the core ID reduction system code achieves a mean error of 0.12 on the test portion of human-annotated similarity corpus, after the similarity scores have been normalized to lie within the [0: 1] range.

To adopt the classifier to our target domain, the parameters learned during the pre-training are subsequently fine-tuned on synthetic similarity data constructed on the basis of the 90k Europarl corpus presented in the 'Data' section. In the absence of similarity

[5] ixa2.si.ehu.es/sts/

scores provided by human experts for the Europarl domain, artificial similarity data was constructed following a naive word replacement strategy. This entails selecting two sentences of similar length – sentence A and sentence B – form the source corpus at random and crossing them over, replacing a random number of words in A with words from B which occur at identical positions in the linear order of the two sentences' surface forms. This process yields two similarity pairs per cross-over operation, as each 'donor' sentence is compared with the outcome of the crossing, with the respective similarity score calculated by dividing the word overlap between the two pair members by the length of the donor sentence. Fine-tuned on the training set of the similarity corpus constructed in this manner, the classifier achieves a mean test error of 0.17, which is comparable to its performance on the pre-training dataset.

As the sentence similarity classifier requires two sequences of discrete word tokens to be provided as its input, it could not be successfully integrated into the proposed system in its current state. The reason behind this incompatibility is that the construction of the ID-reduced output generated by the translator SAE during each training and inference step necessarily involves sampling from its predictive distribution, which constitutes a discrete operation. This, however, interrupts the gradient from the training objective to the generator's parameters, thus making it impossible to train the resulting system via SGD. In light of this, initial steps have been undertaken towards circumventing the non-differentiability issue through the switch to the reinforcement learning (RL) training paradigm. In the context of RL, the sequence prediction task is regarded as sequential sentence generation process, where the generative model is assigned the role of the agent, the previously generated sentence prefix is treated as the state presently occupied by the agent, and the next word token to be generated constitutes the agent's next action. Once a sentence has been fully generated, the agent is assigned a reward based on some specified metric, such as, for instance, a sentence similarity score. An RL algorithm such as REINFORCE [58] can then be used to distribute the reward signal over the individual actions taken by the agent throughout the generation process, thus eliminating the need for the objective functions to be fully differentiable.

In addition to providing access to a more varied set of objective functions, taking the RL route to introducing auxiliary tasks for the ID reduction objective has several dis-

tinct advantages. As the reward score is related to the overall 'goodness' of the complete sequences produced by a generative model, rather than assessing individual words, this approach is inherently more compatible with the system presented in this work as it, too, operates on the sentential level, ultimately attempting to reduce the divergence between distributions of sentence encodings. Moreover, RL helps addressing the exposure bias problem beyond what can be accomplished with scheduled sampling while also facilitating the correct modeling of inter-sentential long distance dependencies in cases where the LSTM cell alone may prove insufficient [48]. Indeed, RL has been successfully combined with adversarial learning for application to natural language generation tasks with encouraging results [60]. In light of these advantages and promising initial studies, this line of inquiry holds great potential for the extension and improvement of the proposed ID reduction engine. Unfortunately, due to time constraints and limited computation resources, this has been left as a direction to be explored in greater detail in the immediate future.

A second auxiliary task which, similarly, is being considered as a guiding criterion for the adversarial training objective is the explicit minimization of the surprisal exhibited by the translator SAE's output. As is the case with sentence similarity estimation, this additional task is meant to complement adversarial loss. It achieves this by explicitly foregrounding surprisal reduction as the learned task, which is only implicitly encouraged through the use of ID-variant corpora as input sources for the adversarial training procedure. Similarly, its calculation also requires for the output of the translator SAE to be fully constructed as a sequence of discrete tokens, before the sentence-wise surprisal score can be calculated, making it inherently incompatible with SGD as the training method. It is, therefore, also regarded as a useful and promising extension to the baseline system, the incorporation of which becomes feasible through the switch to RL.

At its core, the surprisal reduction auxiliary task enforces the lowering of the sentence length normalized surprisal score assigned to the output of the translator SAE during training time. When prefaced with the sigmoid nonlinearity, this ID reduction objective approaches zero as the positive difference in surprisal between the input sentence and the system's output increases. Thus, it is particularly well suited for optimization via loss minimization. The surprisal score for either of the two compared sequences is calculated

with the help of same LM which has been used in the construction of the low-ID and high-ID corpora. As before, this is done by transforming the probability assigned to the sentence continuation at each time step into the corresponding surprisal score in accordance with eq.(1) and subsequently taking the average of the individual word surprisal scores, so as to not penalize longer sentences. As the reduction in surprisal is calculated on the sequence-level, it is fully consistent with the adversarial learning and sentence similarity estimation objectives.

With regard to the particular means by which such RL-based auxiliary objectives can be efficiently incorporated into the architecture proposed herein, one possible approach to follow is outlined in [60]. There, the authors use RL in combination with the GAN framework to train a series of generative models which they then apply to a number of natural language generation tasks. In doing so, they reportedly achieve a better performance than comparable baselines trained via traditional SGD. In this specific scenario, the RL reward is determined by the adversarial learning objective and is the used to update the generator, while the discriminator is trained to distinguish between samples produced by the generator and sentences sampled from a target distribution without recourse to RL. Conceptually, this system bears similarity to the ID reduction engine outlined in this work, in that in both cases learning is guided by a sequence-level objective, with RL being used for this purpose in the former case and informative sentence encodings in the latter.

One way in which RL could therefore be incorporated into the current model is by moving away entirely from relying on sentence encodings as the application point for adversarial learning. Instead, the training signal can be extracted from the decoded sequences produced by the translator generator directly, by comparing them with ground truth low-ID sentences. In doing so, both of the aforementioned auxiliary tasks for similarity enforcement and surprisal reduction can be mixed with the adversarial objective in a linear combination in a straight-forward manner, wighted by their respective importance. Alternatively, the adversarial loss calculation can be left as is, with RL used only to enable the inclusion of the auxiliary tasks. However, as this necessitates two distinct optimization mechanisms to run in parallel during training, a multi-task learning setup where SGD and RL optimization is performed asynchronously on a shared set of parameters – i.e. in a 'hard' parameter sharing configuration – may prove beneficial [49].

While the first method promises to be less computationally expensive, as all objectives are optimized jointly, the use of sentence encodings for the adversarial learning step as envisaged in the second alternative and implemented in the present work may, by itself, be advantageous. Ideally, such sentence embeddings – when specific to the target task and generated by an appropriately trained model – capture the most informative aspects of the sentences from which they are derived. This level of abstraction is not present when adversarial learning is applied to the output sequences directly, which may impact the overall effectiveness and convergence speed of the training procedure negatively. In the absence of positive experimental evidence, however, such assumptions remain speculative. As such, a comparative evaluation of both considered extension types to the extant ID reduction system would be valuable in two ways. On the one hand, either expansion method can be reasonably expected to improve on the performance of the baseline system due to the accessibility of supplementary information relevant to the ID reduction task, namely the sentence similarity and surprisal estimates. On the other hand, a direct comparison would also shed light on how adversarial learning should best be applied to sentences, i.e. whether dense, continuous sentence representations provide a better training signal than sequences of discrete word tokens. An important baseline to consider when evaluating the performance of either extension strategy is an encoder-decoder augmented with both auxiliary objectives but lacking the adversarial component. To objectively asses the contribution of the adversarial training regime to the ID reduction task, regardless of architecture choice, this type of reference for comparison is essential. Exploring either of the two direction for future research, though ideally both, is presently considered as the next step in developing the current system further and into a meaningful direction.

While there undoubtedly exist numerous ways in which the present proposal can be improved upon, especially considering the breakneck pace at which novel GAN variations are introduced and applied to challenging problems, the modifications outlined here can be easily incorporated into the existing system's design. At the moment of writing, the switch to RL constitutes the only remaining challenge, as the pre-trained models for similarity estimation and surprisal scoring have already been obtained and are used in a purely evaluative manner within the presented system. It is my intention to follow up on the outlined proposals for the improvement of the hitherto observed performance in the immediate future and to evaluate their respective contributions in a comprehensive

follow-up study.

6 Conclusion

Having set out to address the challenging, open problem of unsupervised ID reduction, the study documented in this work - while ultimately, falling short of complete success - nonetheless unearthed a plethora of valuable and interesting insights into the examined task as well as the methods chosen to approach it. Among the lessons drawn from the experimental analysis of IDGAN's final performance is the importance of adequate data creation strategies which are well suited to both the trained task and the properties of the trained system; the difficulty involved in balancing multiple objectives differing in their granularity (i.e. sentence-level vs word-level), progression, and stability; and the necessity of the fitness of the chosen adversarial learning objective for its intended functionality, such as the facilitation of ID reduction in a purely data-driven manner.

In light of its contributions, this work aims to serve as a stepping stone for future studies embarking upon the same route. It accomplishes this goal by both offering rough guidance on which architectural decisions and training configurations are more likely to succeed at the ID reduction task, as well as by providing a robust, comprehensive, and modifiable implementation of the IDGAN system. The latter is ostensibly the primary contribution of this work, as it constitutes an effective pipleine for applying adversarial learning to natural language data for the purposes of information density manipulation and is supplemented with numerous methods for linguistic and statistical analysis of corpus data and system outputs.

One such endeavor, which seeks to improve upon the current IDGAN incarnation has been presented in the previous section and is being actively developed at the time of writing. It is the author's hope that this extesion will succeed at utilizing the many insights obtained in the course of the current investigation to surmount some of the open challenges and obtain a better understanding of the means in which unsupervised approaches can be used for the modeling of such complex processes as the reduction of information density.

Appendices

A Hyper-parameter choices

LM hyper-parameter settings for 100k & 90k Europarl corpora	
Parameter	**Value**
Word embedding size	256
Hidden layer size	512
Number of layers	2
Dropout probability	0.5
Sampled softmax samples	25
L2 regularization beta	1e-5
Gradient norm upper bound	10.0
Optimizer	ADAM
Learning rate	0.001
Learning rate annealing steps	2
Learning rate annealing factor	0.9
Unrolled time-steps	35
Batch size	32
Warm-up epochs	6
Stagnant epochs before early stopping	20
Max. training epochs	50

Table 1: LM hyper-parameter choices during corpus construction phase.

SAE hyper-parameter settings for ID-specific corpora	
Parameter	**Value**
Word embedding size	256
Encoder hidden layer size	256
Encoder number of layers	2, unidirectional
Decoder hidden layer size	256
Decoder number of layers	2, unidirectional
Decoder attention size	512
Number of data buckets	8
Dropout probability	0.5
Sampled softmax samples	25
L2 regularization beta	1e-5
Gradient norm upper bound	None
Optimizer	ADAM
Learning rate	0.0001
Learning rate annealing steps	5
Learning rate annealing factor	0.9
Scheduled sampling function	Inverse sigmoid
Scheduled sampling constant	17.48
Batch size	32
Warm-up epochs	10
Stagnant epochs before early stopping	30
Max. training epochs	200

Table 2: SAE hyper-parameter choices during pre-training phase.

References

[1] Martın Abadi et al. “Tensorflow: Large-scale machine learning on heterogeneous distributed systems”. In: *arXiv preprint arXiv:1603.04467* (2016).

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).

[4] Antonio Valerio Miceli Barone. “Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders”. In: *arXiv preprint arXiv:1608.02996* (2016).

[5] Yoshua Bengio et al. “A neural probabilistic language model”. In: *Journal of machine learning research* 3.Feb (2003), pp. 1137–1155.

[6] Denny Britz et al. “Massive exploration of neural machine translation architectures”. In: *arXiv preprint arXiv:1703.03906* (2017).

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.* Vol. 1. IEEE. 2005, pp. 539–546.

[8] Junyoung Chung et al. “Empirical evaluation of gated recurrent neural networks on sequence modeling”. In: *arXiv preprint arXiv:1412.3555* (2014).

[9] Matthew W Crocker, Vera Demberg, and Elke Teich. “Information density and linguistic encoding (IDeaL)”. In: *KI-Künstliche Intelligenz* 30.1 (2016), pp. 77–81.

[10] Andrew M Dai and Quoc V Le. “Semi-supervised sequence learning”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 3079–3087.

[11] Vera Demberg and Frank Keller. “Data from eye-tracking corpora as evidence for theories of syntactic processing complexity”. In: *Cognition* 109.2 (2008), pp. 193–210.

[12] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. "Generative Multi-Adversarial Networks". In: *arXiv preprint arXiv:1611.01673* (2016).

[13] Stefan L Frank. "Uncertainty reduction as a measure of cognitive load in sentence comprehension". In: *Topics in Cognitive Science* 5.3 (2013), pp. 475–494.

[14] Stefan L Frank et al. "The ERP response to the amount of information conveyed by words in sentences". In: *Brain and language* 140 (2015), pp. 1–11.

[15] Richard Futrell, Kyle Mahowald, and Edward Gibson. "Large-scale evidence of dependency length minimization in 37 languages". In: *Proceedings of the National Academy of Sciences* 112.33 (2015), pp. 10336–10341.

[16] Edward Gibson. "The dependency locality theory: A distance-based theory of linguistic complexity". In: *Image, language, brain* (2000), pp. 95–126.

[17] Daniel Gildea and T Florian Jaeger. "Human languages order information efficiently". In: *arXiv preprint arXiv:1510.02823* (2015).

[18] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 249–256.

[19] Ian Goodfellow. "NIPS 2016 tutorial: Generative adversarial networks". In: *arXiv preprint arXiv:1701.00160* (2016).

[20] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680.

[21] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *arXiv preprint arXiv:1704.00028* (2017).

[22] John Hale. "A probabilistic Earley parser as a psycholinguistic model". In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics. 2001, pp. 1–8.

[23] John Hale. "Information-theoretical complexity metrics". In: (2016).

[24] John Hale et al. "Modeling fMRI time courses with linguistic structure at various grain sizes". In: *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*. 2015, pp. 89–97.

[25] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[26] Jonathan Ho and Stefano Ermon. “Generative adversarial imitation learning”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 4565–4573.

[27] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[28] Matthew Honnibal. *spaCy: Industrial strength NLP with Python and Cython*. 2015. URL: `spacy.io`.

[29] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International Conference on Machine Learning*. 2015, pp. 448–456.

[30] Ozan Irsoy and Claire Cardie. “Deep recursive neural networks for compositionality in language”. In: *Advances in neural information processing systems*. 2014, pp. 2096–2104.

[31] T Florian Jaeger and Roger P Levy. “Speakers optimize information density through syntactic reduction”. In: *Advances in neural information processing systems*. 2007, pp. 849–856.

[32] Sébastien Jean et al. “On using very large target vocabulary for neural machine translation”. In: *arXiv preprint arXiv:1412.2007* (2014).

[33] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. “An empirical exploration of recurrent network architectures”. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 2015, pp. 2342–2350.

[34] Rafal Jozefowicz et al. “Exploring the limits of language modeling”. In: *arXiv preprint arXiv:1602.02410* (2016).

[35] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. “A convolutional neural network for modelling sentences”. In: *arXiv preprint arXiv:1404.2188* (2014).

[36] Diederik Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[37] Roger Levy. "Expectation-based syntactic comprehension". In: *Cognition* 106.3 (2008), pp. 1126–1177.

[38] Roger Levy. "Memory and surprisal in human sentence comprehension". In: *Sentence processing* 78 (2013).

[39] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

[40] Alireza Makhzani et al. "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644* (2015).

[41] Xudong Mao et al. "Least squares generative adversarial networks". In: *arXiv preprint ArXiv:1611.04076* (2016).

[42] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[43] Jonas Mueller and Aditya Thyagarajan. "Siamese Recurrent Architectures for Learning Sentence Similarity." In: *AAAI*. 2016, pp. 2786–2792.

[44] Youssef Oualil and Dietrich Klakow. "A Neural Network approach for mixing language models". In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE. 2017, pp. 5710–5714.

[45] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.

[46] Ofir Press and Lior Wolf. "Using the output embedding to improve language models". In: *arXiv preprint arXiv:1608.05859* (2016).

[47] Sai Rajeswar et al. "Adversarial Generation of Natural Language". In: *arXiv preprint arXiv:1705.10929* (2017).

[48] Marc'Aurelio Ranzato et al. "Sequence level training with recurrent neural networks". In: *arXiv preprint arXiv:1511.06732* (2015).

[49] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[50] Tim Salimans et al. “Improved techniques for training gans”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2226–2234.

[51] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural networks* 61 (2015), pp. 85–117.

[52] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural machine translation of rare words with subword units”. In: *arXiv preprint arXiv:1508.07909* (2015).

[53] Binoy Shah and Howie Choset. “Survey on urban search and rescue robots”. In: *Journal of the Robotics Society of Japan* 22.5 (2004), pp. 582–586.

[54] Claude Elwood Shannon. “A mathematical theory of communication”. In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.

[55] Nathaniel J Smith and Roger Levy. “The effect of word predictability on reading time is logarithmic”. In: *Cognition* 128.3 (2013), pp. 302–319.

[56] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. “Practical bayesian optimization of machine learning algorithms”. In: *Advances in neural information processing systems*. 2012, pp. 2951–2959.

[57] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[58] Ronald J Williams. “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. In: *Machine learning* 8.3-4 (1992), pp. 229–256.

[59] Zichao Yang et al. “Hierarchical Attention Networks for Document Classification.” In: 2016.

[60] Lantao Yu et al. “SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient.” In: 2017.

[61] Yizhe Zhang and Zhe Gan. “Generating Text via Adversarial Training”. 2016.

[62] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *arXiv preprint arXiv:1703.10593* (2017).