Syntax and Geometry of Information

Raphael Bailly, Kata Gábor, Laurent Leblond

Syntax and Geometry of Information

Raphaël Bailly
SAMM, EA 4543, FP2M 2036 CNRS
Université Paris 1
rbailly@univ-paris1.fr

Laurent Leblond
Stellantis
laurent.leblond1@stellantis.com

Kata Gábor
ERTIM, EA 2520
INALCO
kata.gabor@inalco.fr

Abstract

This paper presents an information-theoretical model of syntactic generalization. We study syntactic generalization from the perspective of the capacity to disentangle semantic and structural information, emulating the human capacity to assign a grammaticality judgment to semantically nonsensical sentences. In order to isolate the structure, we propose to represent the probability distribution behind a corpus as the product of the probability of a semantic context and the probability of a structure, the latter being independent of the former. We further elaborate the notion of abstraction as a relaxation of the property of independence. It is based on the measure of structural and contextual information for a given representation. We test abstraction as an optimization objective on the task of inducing syntactic categories from natural language data and show that it significantly outperforms alternative methods. Furthermore, we find that when syntax-unaware optimization objectives succeed in the task, their success is mainly due to an implicit disentanglement process rather than to the model structure. On the other hand, syntactic categories can be deduced in a principled way from the independence between structure and context.

1 Introduction

In the context of both human learning and statistical machine learning, what distinguishes generalization from memorization is the process of deliberately ignoring a part of the input information. In machine learning, the generalization will be guided towards some specific direction by the learning hypothesis: the model structure and the choice of regularization impacts the nature of the information loss. We can talk about syntactic generalization from textual data when the information pertaining to sentence structure tend to be preserved by the model and the information pertaining to other aspects of the text tend to be ignored.

Syntactic generalization is of great interest because the human capacity to assign an abstract structure to utterances is a prerequisite to creatively combine constituents and understand novel sentences (Frege, 1892). Knowledge of syntax can boost the robustness of NLP applications with respect to unseen data, in particular when there is a distribution shift (He et al., 2020; Wu et al., 2019). In a broader perspective, understanding syntactic generalization informs the discussion on the learnability of syntax from unlabelled text without any built-in grammatical knowledge or inductive bias (Gold, 1967; Clark and Lappin, 2010; Bailly and Gábor, 2020). Finally, studying syntactic generalization in large language models (LLMs) sheds light on whether and to what extent these models emulate human functioning with respect to linguistic competence.

The prevailing formalization of syntax is by means of algebraic compositional rules operating on a finite set of discrete categories (parts of speech). Language models can acquire syntactic knowledge when they are given specific supervision or bias (Dyer et al., 2016; Shen et al., 2020; Sartran et al., 2022). Whether unsupervised settings can lead to syntactic generalization and under which conditions is still unknown. Current LLMs use distributed representations that cannot be unequivocally mapped to a set of categories, let alone syntactically meaningful categories. The question whether their representations encode syntactic information and how to uncover it is actively investigated today (Hu et al., 2020; Marvin and Linzen, 2018). The majority of works in the topic of syntactic generalization in language models adopt an empirical approach, such as probing or analysis of a model with a comparison to actual or expected human performance on linguistically motivated tasks. In contrast, we present a theoretical approach to formalize syntactic generalization in an information theoretical

framework.

Statistical learning can be formulated as the minimization of KL-divergence - a measure of information loss - subject to constraints. The constraints on model expressivity ensure that generalization takes place by eliminating the information resulting from sampling noise. We claim that the training objective of maximum likelihood estimation by nature does not incentivize models to syntactic generalization. In the case of syntactic generalization, the information loss needs to be directed to non-structural information, which is only remotely related to the elimination of sampling noise.
First, a corpus is not randomly sampled from the set of grammatical sentences. Word co-occurrences in a corpus are indeed influenced by different factors such as semantics and pragmatics. The process of abstracting away from these factors is arguably different from the concept of generalization in machine learning, as the acquisition of syntactic knowledge always involves a shift of distribution (Hupkes et al., 2022). Second, the target of generalization is the capacity to recognize the *set* of grammatical sentences: well-formedness is inherently a binary notion rather than a probabilistic one. These considerations motivate our proposition to decompose a corpus distribution as a factor of semantic/pragmatic context and a factor of structure representing well-formedness. In what follows, we reinterpret syntactic generalization based on the separation of structural and semantic information, and we show that our approach outperforms concurrent methods on unsupervised POS induction. We also define the notion of abstraction, an optimization objective specifically conceived for disentangling semantic information and syntactic well-formedness.

1.1 Related work

Generative linguists agree on the nativist argument that learners cannot converge on the same syntax unless some of their linguistic knowledge is innate (Baker, 1979; Chomsky, 1965, 1975), which makes the complete unsupervised learning of syntax impossible. Therefore, theoretical linguistics demonstrated little interest in machine learning and the interaction between the two fields is limited (Lappin and Shieber, 2007; Linzen and Baroni, 2021).[1]

[1]The nativist claim has however been subjected to criticism from other fields of research (Hsu and Chater (2010); Yang and Piantadosi (2022).

With the recent advent of large language models (Devlin et al., 2019; Peters et al., 2018; Radford et al., 2019) it has become relevant to test their linguistic competence (Linzen et al., 2016; Belinkov and Glass, 2019; Baroni, 2019). Researchers in NLP thus turned to linguistic theory to create probing tasks (Alain and Bengio, 2017; Giulianelli et al., 2018) or test sets targeted at specific linguistic knowledge (Linzen et al., 2016). Linguistic challenges like long-distance agreement (Linzen et al., 2016), hierarchical syntax (Lin et al., 2019; Dyer et al., 2016; Conneau et al., 2018; Hupkes et al., 2018), parts of speech (Saphra and Lopez, 2018; Kim and Smolensky, 2021), or morphology (Belinkov et al., 2017; Peters et al., 2018) have been applied to probe the latest language models with contrasting results. Recently, probing classifiers have also been subjected to methodological criticism. Models can succeed on some test tasks by learning shallow heuristics (McCoy et al., 2019; Poliak et al., 2018). It was also argued that the presence of sufficient information to learn a given task does not entail alone that models rely on it Ravichander et al. (2021); Hewitt and Liang (2019); Xu et al. (2020).
On rarer occasions, studies aimed to test the capacity of language models to predict grammaticality judgments. Out of distribution testing systematically shows that the performance drops when the test data contains natural or artificial examples which are deliberately different from the training examples (Lake and Baroni, 2017; Marvin and Linzen, 2018; Chowdhury and Zamparelli, 2018; van Schijndel et al., 2019; Maudslay and Cotterell, 2021).

Another branch of model analysis and interpretation studies are concerned with the nature of the generalization that takes place, with a particular accent on the notion of *compositionality* (Loula et al., 2018; Baroni, 2019; Valvoda et al., 2022). Among others, Fodor and Lepore (2002) and Kottur et al. (2017) claim that syntactic compositionality (Chomsky, 1957, 1965) is a prerequisite to learn to generalize to complex unseen input. In empirical studies, Gulordava et al. (2018) and Lakretz et al. (2019) report a lack of compositionality in the models they analyse, despite their impressive performance. In contrast, Bastings et al. (2018) and Valvoda et al. (2022) find that some compositional relations can be learned by neural sequence-to-sequence models. Chaabouni et al. (2020) argue

that there is no correlation between the compositionality of an emergent language and its ability to generalize.

The problem of conflation between semantic and syntactic information in language models has been identified (Maudslay and Cotterell, 2021) as a factor hindering syntactic generalization. A new line of research is concerned with disentangling syntactic and semantic information in representations (Felhi et al., 2020; Huang et al., 2021) by adversarial training or syntactic supervision. In order to incite syntactic generalization in models, Shen et al. (2020) and (Dyer et al., 2016) propose to integrate explicit syntactic information for language modelling. Hu et al. (2020) show that there is a trade-off between the general language modelling objective and syntax-specific performance.

Some recent work relies on information theory to improve our understanding of the syntactic knowledge in LMs. Pimentel et al. (2020) reformulates probing as approximating mutual information between a linguistic property and a contextual representation. Subsequently, Pimentel and Cotterell (2021) introduced Bayesian Mutual Information, a definition that allows information gain through processing. Voita and Titov (2020) use Minimum Description Length to measure regularity in LM representations with respect to the labels to be predicted in a linguistic probe. Our work builds on the propositions formulated in Bailly and Gábor (2020) who address the problem of the learnability of grammar by separating syntactic and semantic information in a corpus.

2 Syntactic Representation

2.1 Autonomy of Syntax

The concept of generalization we introduce is based on the autonomy of syntax (Chomsky, 1957, 1982; Adger, 2018) reinterpreted in terms of statistical independence. In the process of linguistic generalization, learners need to abstract away from semantic, pragmatic and idiosyncratic lexical information in the input they are exposed to. With a string prediction task and likelihood maximization as a training objective, models have no incentive to abstract away from these features. One can expect a statistical learner to ignore sampling noise, but the above features are relevant to learn the distribution behind a corpus. This insight motivates our proposition of *statistical abstraction*, a training objective that focuses on certain aspects of the input while deliberately ignoring others.

We want our learner to concentrate on the structure and ignore the factors we call *context*, i.e. all the aspects that are unrelated to well-formedness. We do so by creating two representations of the input: one of them structured, the other having structural information removed but co-occurrence relations conserved.

Let us consider a small artificial example for illustration. Our observation is a corpus with the two sentences below:

cats eat mice
men build houses

A valid syntactic generalization would recognize the sentence

cats build mice

as grammatical. In order to do so, we consider

$$p(\text{cats eat mice})$$

as a factor of the probability of the co-occurrence of its words in the same *context* :

$$p(\{\text{cats, eat, mice}\})$$

and a factor of the probability of the words to appear in a given *structure* :

$$p(\text{cats eat mice}|\{\text{cats, eat, mice}\})$$

A syntactic representation with a desirable degree of generalization would identify the distributional classes {cats, men}, {build, eat}, {mice, houses}.

This set of distributional classes can be seen as a function f that associates a word (e.g *cats*) with its class (*{cats, men}*). Our goal is to study the properties of such a function so that it can be considered as achieving syntactic generalization, for instance:

$$p(\text{cats eat mice}|\{\text{cats, eat, mice}\})$$

can be deduced from

$$p(\text{f(cats) f(eat) f(mice)}|\{\text{f(cats), f(eat), f(mice)}\})$$

2.2 Properties of a Syntactic Partition

We define the probability distribution that predicts the grammaticality of sequences, learned from observation. In order to do so, we first define a partition of words into abstract categories. This mapping, together with the category sequences found in the corpus, will allow us to induce the grammar.

Behind the corpus data there is a probability distribution $p(w_1 w_2 \dots w_n)$. This distribution can be written as a product of two factors. First, the unstructured data, i.e. the probability of the elements

of the vocabulary to occur in the same sequence without considering their order. Second, the probability of these elements to be observed in a particular structure. The *contextual* information is related to the former, and the *structural* information to the latter.

Let us see a probabilistic interpretation. Let A be the vocabulary, one defines the set $A^+ = A^*\backslash\varepsilon$ where ε is the empty sequence. $\boldsymbol{w} = w_1 \dots w_n \in A^n$ a sequence (of words) of length n and $p(\{w_1, \dots, w_n\})$ the probability of observing these elements in the same sequence, in any order. A trivial decomposition of $p(w_1 w_2 w_3)$ would be

$$p(\{w_1, w_2, w_3\})p(w_1 w_2 w_3 | \{w_1, w_2, w_3\})$$

However, we want structural information to be independent of the context. The decomposition above does not suppose the autonomy of structure. We propose to transform the above distribution with a mapping f, which will induce a partition over the elements of the vocabulary. In what follows, we examine which properties of this mapping will ensure that the categories of the resulting partition do not contain contextual information, while still preserving the information necessary to predict grammaticality.

$\boldsymbol{w} = w_1 \dots w_n \in A^n$
$\lvert\boldsymbol{w}\rvert =$ length of $\boldsymbol{w}$
$f(\boldsymbol{w}) = f(w_1) \dots f(w_n)$
$f[\boldsymbol{w}] = \{\boldsymbol{w'} \in A^+ \mid f(\boldsymbol{w'}) = f(\boldsymbol{w})\}$
for $\sigma \in \mathfrak{S}_n$, $\sigma(\boldsymbol{w}) = w_{\sigma(1)} \dots w_{\sigma(n)}$
$\langle\!\langle W \rangle\!\rangle = \cup_{\sigma \in \mathfrak{S}_n, \boldsymbol{w} \in W} \{\sigma(\boldsymbol{w})\}$
$\mu(\boldsymbol{w}) = \text{card}(\{\sigma \in \mathfrak{S}_n \mid \boldsymbol{w} = \sigma(\boldsymbol{w})\})$

Table 1: Notations

Let A be the vocabulary, one defines the set $A^+ = A^*\backslash\varepsilon$ where ε is the empty sequence.

Let $f(\boldsymbol{w})$ denote the sequence of categories resulting from the mapping of a word sequence, and $f[\boldsymbol{w}]$ the set of sequences that map to $f(\boldsymbol{w})$. W denotes a set of sequences $\boldsymbol{w}$. In the case of a singleton we will denote $\langle\!\langle \boldsymbol{w} \rangle\!\rangle = \langle\!\langle \{\boldsymbol{w}\} \rangle\!\rangle$. The contextual information will be modeled through the probability $p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$, where one can see the object $\langle\!\langle \boldsymbol{w} \rangle\!\rangle$ as a bag of words, from which the information of the structure (order) has been erased.

A syntactically relevant representation needs to meet two criteria: it has to allow to recover the structure, i.e. the ordering of the bag of words, and it needs to be independent of contextual information. The first criterion is defined below as factorization, the second as minimality.

Factorization. One will say that a mapping f factorises a distribution p if the order of a bag-of-words $\{w_i\}$ drawn from p can be entirely deduced from the knowledge of the corresponding categories.

Definition 1. Let p be a distribution over A^+, and $f : A \mapsto B$ be a mapping. The distribution p is factorised by f if there exists a mapping $\lambda_f(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$ such that $\forall \boldsymbol{w} \in A^+$

$$p(\boldsymbol{w} \mid \langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \lambda_f(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)\, p(f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)$$

in that case, one has $\lambda_f(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \frac{\mu(\boldsymbol{w})}{\mu(f(\boldsymbol{w}))}$.

In the case where f factorises p, one will say that context and structure are independent conditionally to f.

Independence. As the property of factorization does not guarantee the complete independence of structure and context (for instance the identity always factorises p), we need to limit the information carried by f to its minimal value in order to reach this independence. From $f[\boldsymbol{w}]$ one can deduce, at the minimum, the length of $\boldsymbol{w}$. The purpose of minimality is to ensure that knowing $f[\boldsymbol{w}]$ provides no further information for finding $\boldsymbol{w}$:

Definition 2. Let p be a distribution over A^+ and let $f : A \mapsto B$ be a mapping. We will say that f is (information)-minimal for p if

$$\forall \boldsymbol{w} \in A^+, p(\boldsymbol{w} \mid f[\boldsymbol{w}]) = p(\boldsymbol{w} \mid A^{|\boldsymbol{w}|})$$

We will say that context and structure are independent in p if there exists an information-minimal factorization of p.

2.3 Induced grammar

From a probability distribution p and a mapping f, it is possible to induce a syntax based on the observed patterns: a sequence is structurally correct if its pattern corresponds to an observed pattern.

Definition 3. Let p be a distribution over A^+ and let $f : A \mapsto B$ be a mapping. One denotes the syntax induced by p and f by

$$\boldsymbol{w} \in \mathcal{G}(p, f) \Leftrightarrow p(f[\boldsymbol{w}]) > 0$$

One has for instance $\mathcal{G}(p, id) = \mathit{supp}(p)$: this representation is a memorization with no generalization.

Minimal syntax. A syntax induced by minimal factorization of p will be called minimal syntax. The set of all minimal syntaxes will be denoted $G^*(p)$.

It can be shown that the intersection of all minimal syntaxes of p is a minimal syntax of p:

$$\mathcal{G}^*(p) = \cap_{f \in G^*(p)} \mathcal{G}(p, f) \in G^*(p)$$

Hence, if the independence between context and structure holds, there exists a canonical way to define the set of well-structured sequences which is different from the support of p.

Example 1. Let us consider the first example above: let p be the distribution defined by

$$p(\textit{cats eat mice}) = \tfrac{1}{2}$$
$$p(\textit{men build houses}) = \tfrac{1}{2}$$

then the mapping f defined by

$$f(\textit{cats}) = f(\textit{men}) = b_0$$
$$f(\textit{eat}) = f(\textit{build}) = b_1$$
$$f(\textit{mice}) = f(\textit{houses}) = b_2$$

is a minimal factorization of p. The minimal syntax $\mathcal{G}^*(p)$ is the set

cats eat mice	cats eat houses
cats build mice	cats build houses
men eat mice	men eat houses
men build mice	men build houses

3 Geometry of Information

Using information theoretical tools, we transform the criteria above into metrics and define an information space which allows to track the amount of contextual and structural information in a partition, as well as the direction of generalization during a training process.

The concept of minimal factorization provides the formal definition of minimal syntax; however, the conditions of factorization (Definition 1) and minimality (Definition 2) are restrictive. In natural language corpora, a perfect independence between semantic context and grammaticality cannot be expected. Syntax and semantics do interface in natural language, semantic acceptability interacts with grammaticality and depending on how one deals with this interface, either the assumption of perfect independence or the precise retrieval of the distribution underlying the corpus may not be met. This motivates our methodology for relaxing both conditions in a way that gives an equivalent but quantifiable formulation for each criterion in terms of information. We thus provide a method to measure the amount of structural information present in a partition, hence relaxing the factorization criterion. We also define contextual information, which relaxes the minimality requirement.

3.1 Structural information

Let

$$H(p \parallel q) = -\sum_{\boldsymbol{w} \in A^+} p(\boldsymbol{w}) \log(q(\boldsymbol{w}))$$

be the cross entropy of the distribution q with respect to the distribution p.

For a distribution p over A^+, we will consider the distance (in terms of cross-entropy) between p and the class of factorised distributions.

Definition 4. Let p be a distribution over A^+ and let f be a mapping. One denotes:

$$\mathcal{F}_f = \{q \mid q \text{ is factorised by } f\}$$

and one defines the projection of p conditionally to f by

$$p_{|f} = \arg\min_{q \in \mathcal{F}_f} H(p \parallel q)$$

The structural information of f with respect to p is given by

$$i_s(p \parallel f) = H(p \parallel p_{|z}) - H(p \parallel p_{|f})$$

where z is the null mapping.

The set $\mathcal{F}_f$ represents the set of distributions for which the knowledge of f is sufficient to recover the order of a sequence. The structural information is minimal for z, and maximal for the identity (see Appendix):

$$i_s(p \parallel z) = 0 \leq i_s(p \parallel f) \leq i_s(p \parallel id)$$

The link between structural information and factorization is given by:

Lemma 1. Let p be a distribution over A^+ and let $f : A \mapsto B$ be a mapping. One has

$$i_s(p \parallel f) \text{ is maximal } \Leftrightarrow f \text{ factorises } p$$

3.2 Contextual information

An optimal syntactic representation is one that fulfills the independence requirement: the probability of a sequence of categories does not provide information about which actual words are likely to

appear in the sentence. The contextual information will measure the amount of lexical or semantic information that is present in a representation.

Let

$$H(p) = H(p \parallel p)$$

be the Shannon entropy. Let p be a distribution over A^+ and let $f : A \mapsto B$ be a mapping. One will denote $p \circ f^{-1}$ the distribution on B^+ induced by f. One has $p \circ f^{-1}(f(w)) = p(f[w])$.

Definition 5. The contextual information of f with respect to p is given by

$$i_c(p \parallel f) = H(p \circ f^{-1}) - H(p \circ z^{-1})$$

where z is the null mapping.

From standard properties of Shannon entropy, $i_c(p \parallel f)$ is minimal for z, and maximal for the identity (see Appendix):

$$i_c(p \parallel z) = 0 \leq i_c(p \parallel f) \leq i_c(p \parallel id)$$

The maximum value of $i_c(p \parallel f)$ is reached for $H(p \circ f^{-1}) = H(p)$.

The link between contextual information and information-minimality is given by:

Lemma 2.

$$i_c(p \parallel f) = 0 \Leftrightarrow f \text{ is minimal for } p$$

3.3 Representation of a mapping in the information space

Let us now consider how to represent geomtrically the two types of information in a partition. For a given distribution p, any mapping f will be represented in $\mathbb{R}^2$ by its coordinates

$$x_f = i_s(p \parallel f), \quad y_f = i_c(p \parallel f)$$

Example 2. Let us consider the same distribution as in Example 1. Fig. 1 represents all possible mappings g by the point with coordinates $(i_s(p \parallel g), i_c(p \parallel g))$.

Details are in the Appendix. One can check that the minimal factorization is in $(1, 0)$, and the second closest mapping to a minimal factorization is $\{cats, mice, men, houses\}\{eat, build\}$.

4 Abstraction

Abstraction relaxes the definition of a minimal factorization of p in terms of a solution to an optimization problem. For a given probability distribution p and a mapping f, the abstraction measures the distance between f and the position of a minimal factorization of p in the information space:

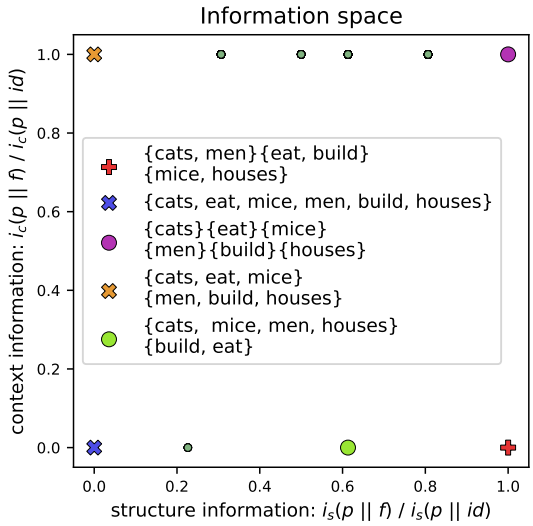

Figure 1: Information space: normalised representation of partitions for the distribution p in Example 1.

Definition 6. Let p be a distribution over A^+ and let $f : A \mapsto B$ be a mapping. Let d be a distance on $\mathbb{R}^2$. Let $t_f = (i_s(p \parallel f), i_c(p \parallel f))$ and $t^* = (i_s(p \parallel id), 0))$. The abstraction (w.r.t. d) is defined as

$$\alpha_d(p \parallel f) = e^{-d(t_f, t^*)}$$

One has $\alpha_d(p \parallel f) \leq 1$, with the maximum value reached iff f is a minimal factorization of p. For a mapping f, maximizing abstraction can be considered as a relaxation of the property of being a minimal factorization.

4.1 Minimal Syntax Identification

We prove here that abstraction can be used to identify the set of minimal syntaxes of p from a sample.

Consistency of the plug-in estimator of abstraction. In the case where the set of of possible sequences is infinite, it is not possible to ensure a convergence rate of the abstraction (cf.(Antos and Kontoyiannis, 2001)). Nevertheless, it is possible to show the following consistency result:

Proposition 1. Let p be a distribution over A^+, and let d be a distance on $\mathbb{R}^2$. Let $\hat{p}_N$ be the empirical distribution derived from an i.i.d. sample of size N drawn from p.

The plug-in estimator for the abstraction $\alpha_d(p||f)$ is consistent:

$$\alpha_d(\hat{p}_N \parallel f) \overset{N \to \infty}{\longrightarrow} \alpha_d(p \parallel f) \; a.s.$$

As a consequence, when the vocabulary A is finite, abstraction can be used to isolate the set of minimal factorizations of p.

Corollary 1. Let p be a distribution over A^+, with $|A| < \infty$. Let d be a distance measure on $\mathbb{R}^2$. Let $\hat{p}_N$ be the empirical distribution derived from an i.i.d. sample of size N drawn from p.

Then one has:

$$\lim_{N\to\infty} \boldsymbol{P}[\mathcal{G}(p, f_d^*(\hat{p}_N)) \in G^*(p)] = 1$$

where $f_d^*(\hat{p}_N)$ maximizes abstraction for $\hat{p}_N$.

5 Experiments

We test abstraction as an optimization objective for learning syntactic representations, when the representation takes the form of a mapping into discrete syntactic categories. The results are evaluated on an unsupervised POS induction task. While our understanding of a syntactic category may not perfectly overlap with actual parts of speech (the latter being defined on the basis of a mixture of criteria instead of pure syntax, and are usually more coarse-grained than real distributional categories), this task will allow a good comparison with concurrent models on a gold standard.

In NLP, part-of-speech categories are usually a part of a probabilistic model; typically a parameter which will be tuned during learning. For instance, if the model is an HMM, its hidden states correspond to POS categories. If the model is a PCFG, categories will correspond to non-terminals. We call this approach - when POS categories are deduced from a given model structure as a parameter - the model-specific approach. In the experiments, we compare the model-specific approach with our hypothesis: that POS categories can be deduced from the independence of structure and context. We consider the task of unsupervised POS induction, and compare the accuracy of the abstraction maximization criterion with model-specific cross-entropy minimization.

The corpus we use comes from Wikipedia in simplified English, contains 430k sentences, 8M tokens, and was POS tagged by the Stanford POS tagger (Toutanova et al., 2003). To create the target partition, words (a vocabulary of 6044 elements) were assigned to their most frequent POS. There are 36 POS categories.

5.1 The target partition in the information space

We created (Fig. 2) the information space for the Wikipedia corpus with the coordinates indicating

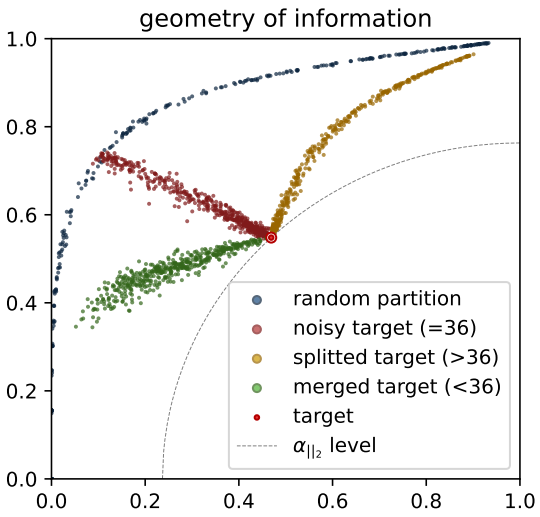

Figure 2: Information space for the Wikipedia corpus. Neighbourhood of the target partition, with and without changing the number of categories.

structural and contextual information. We represented the target partition (36 categories, correct mapping), and located randomly generated modifications of this partition obtained by changing 1) the assignment of words to the target POS categories (in red) and 2) the number of categories between 2 and 2000 (partitions with > 36 categories are in yellow, partitions with < 36 categories in green) by merging or splitting existing categories. First, it can be observed that any random modification of the target partition (whether it increases or decreases the information) comes at the expense of the abstraction objective. This distinctive position of the syntactic partition could not be visualized in one dimension, suggesting the relevance of the coordinates in the information space in identifying it.

Second, with a strict constraint on the number of categories, the representation of the noisy target (in red) indicates a negative correlation between contextual information and structural information: a trade-off induced by the limitation of information capacity. The choice of normalised $d_{||_2}$ distance for abstraction is driven by the shape of random partitions in the information space (in blue).

5.2 POS induction

We compare abstraction and likelihood maximization as training objectives for unsupervised POS induction. The most efficient POS induction methods at present are mainly – if not exclusively –

based on models derived from HMM (Brown et al., 1992; Merialdo, 1994; Lin et al., 2015; Stratos et al., 2016; Tran et al., 2016; He et al., 2018). We experiment with different variations of the model by Brown et al. (1992), because the method is purely distributional, involves discrete embeddings and is still competitive (cf. (Stratos et al., 2016; Christodoulopoulos et al., 2010)).

As we cannot perform a brute-force search for the best possible partitions for our criteria, we replaced it by a local measure of the performance : for every single word, provided that all other words are correctly classified, we checked whether the criterion would attribute the correct POS category. Accuracy indicates the rate of correctly classified words.

Tested models We will call plain model the general form of a distribution p factorised by a mapping f:

$$p(f[\boldsymbol{w}])p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)\frac{\mu(\boldsymbol{w})}{\mu(f(\boldsymbol{w}))}$$

We can add model-specific constraints:
(MK): Markov constraint for

$$p(f[\boldsymbol{w}]) = p(f(w_1))\prod_{i=2}^{n} p(f(w_i) \mid f(w_{i-1}))$$

(CI): contextual independence constraint for

$$p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) = \prod_{i=1}^{n} p(w_i \mid f(w_i))\frac{\mu(f(\boldsymbol{w}))}{\mu(\boldsymbol{w})}$$

We will consider the normalised $\alpha_{||_2}$ abstraction maximization objective, and the likelihood maximization objective (with a constraint on the number of categories) for the plain model alone, with contextual independence (CI) constraint, with Markov (MK) constraint, or with both constraints (MK) + (CI) (Brown clustering criterion).

The results are shown in Figure 3. They indicate that the abstraction criterion significantly outperforms likelihood maximization for any model considered. This reinforces our hypothesis that syntactic categories emerge naturally from the criterion of independence between structure and context, without any assumption about the structure of the model.

The second important finding concerns the phenomenon we call **implicit disentanglement**. By

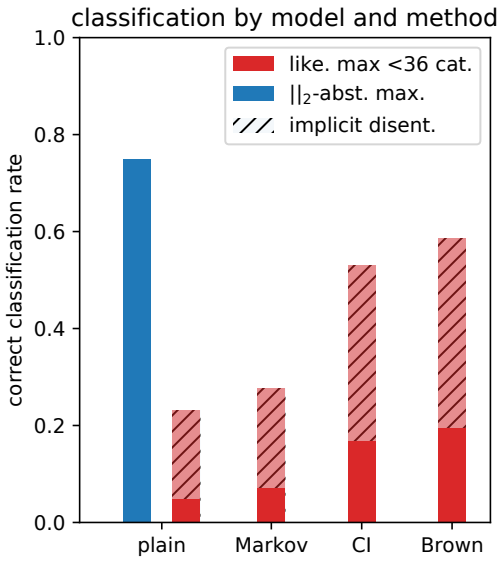

Figure 3: Unsupervised POS induction: abstraction vs likelihood maximization

definition, if we estimate the parameters of a distribution $q_{|f}$ with likelihood maximization, we maximize structural information (i.e. the partition f tends towards the right-most solutions in the information space). However, contextual information will still be present. Syntactic generalization may occur when the encoding capacity of the model is bounded (e.g. by limiting the number of categories), inducing a trade-off between structural and contextual information.
A way to estimate the role of implicit disentanglement is to consider the confusion matrix of correctly classified or misclassified words for abstraction maximization classifier (a) and a likelihood maximization classifier (b), and decompose (b) into a convex combination of (a) and an independent classification process (c).

With the confusion matrix for Brown clustering criterion:

$$M = \begin{array}{c} \\ \overline{a} \\ a \end{array} \begin{array}{c} \begin{array}{cc} \overline{b} & b \end{array} \\ \begin{pmatrix} 0.202 & 0.049 \\ 0.212 & 0.537 \end{pmatrix} \end{array}$$

one obtains that a proportion $F = 0.667$ of the correct classification of (b) is imputable to (a), and at most 33.3% of correct classification by (b) can be considered as independent from implicit disentanglement. This factor F is known in literature as certainty factor (see (Tan et al., 2002), Appendix for details)

The accuracy of maximum likelihood with the plain model (no constraint) is a good example of

implicit disentanglement : it can only be the result of the limitation on the number of categories. The hatched part in Figure 3 represents the fraction of correct classification due to implicit disentanglement in max-likelihood classifiers.
These results indicate that the impact of model structure in the ability to infer syntactic categories (and, more broadly, in syntactic generalization capacity) is over-estimated: parameter tuning seems far less efficient than the application of the principle of independence between context and structure.

6 Conclusion

As to our current knowledge, language models do not have a convincing performance on modelling grammaticality: despite their impressive results on downstream tasks, they are not good at syntactic generalization unless syntactic knowledge is somehow injected in the system. Moreover, there is a trade-off in large LMs between syntactic generalization and language modelling performance.

We suggest a measurable interpretation of syntactic generalization and show results that align with the observations reported by many authors: training on a natural language corpus (e.g. using language models) results in memorization of semantics and entanglement with syntactic information. This motivates our proposition of abstraction, a new training objective for syntactic generalization without supervision. We prove the statistical consistency of abstraction in the task of grammar identification. Empirical results on an unsupervised POS induction task show that abstraction considerably outperforms concurrent models trained with a likelihood estimation objective, without making any assumptions about the structure of the model.

7 Limitations

The contribution of this paper is mainly theoretical. Like most of the POS identification algorithms, the optimization of a criterion among the space of all partitions requires the use of heuristics, and finding the optimum is never guaranteed. Additional work is required before a generalization model that is efficient in practice can be obtained.

8 Acknowledgement

We would like to thank Guillaume Wisniewski and the anonymous reviewers for their valuable comments.

References

David Adger. 2018. The autonomy of syntax. In Norbert Hornstein, Howard Lasnik, Pritty Patel-Grosz, and Charles Yang, editors, Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics, pages 153–176. De Gruyter Mouton.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. ArXiv, abs/1610.01644.

András Antos and Ioannis Kontoyiannis. 2001. Convergence properties of functional estimates for discrete distributions. Random Structures & Algorithms, 19(3-4):163–193.

Raphaël Bailly and Kata Gábor. 2020. Emergence of syntax needs minimal supervision. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics.

C. L. Baker. 1979. Syntactic theory and the projection problem. Linguistic Inquiry, 10:533–581.

Marco Baroni. 2019. Linguistic generalization and compositionality in modern artificial neural networks. CoRR, abs/1904.00157.

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 47–55, Brussels, Belgium. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 861–872.

Yonatan Belinkov and James R. Glass. 2019. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7:49–72.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. Computational Linguistics, 18(4):467–480.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4427–4442, Online. Association for Computational Linguistics.

Noam Chomsky. 1957. Syntactic Structures. Mouton, Berlin, Germany.

Noam Chomsky. 1965. Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.

Noam Chomsky. 1975. Reflections on language. New York: Pantheon Books.

Noam Chomsky. 1982. Some concepts and consequences of the theory of government and binding. MIT Press, Cambridge, Mass.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In Proceedings of the 27th International Conference on Computational Linguistics, pages 133–144.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 575–584, Cambridge, MA. Association for Computational Linguistics.

Alexander Clark and Shalom Lappin. 2010. Unsupervised learning and grammar induction. In Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwell, Oxford.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 2126–2136.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

G. Felhi, Joseph Le Roux, and Djamé Seddah. 2020. Disentangling semantics in language through vaes and a certain architectural choice. ArXiv, abs/2012.13031.

Jerry A. Fodor and Ernest Lepore. 2002. Compositionality Papers. Oxford University Press UK.

Gottlob Frege. 1892. Über Sinn und Bedeitung. Zeitschrift für Philosophie und philosophische Kritik, 100:25–50.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP, pages 240–248.

E. Mark Gold. 1967. Language identification in the limit. Information and control, 10:5:447–474.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1195–1205.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Unsupervised learning of syntactic structure with invertible neural projections. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1292–1302, Brussels, Belgium. Association for Computational Linguistics.

Q. He, H. Wang, and Y. Zhang. 2020. Enhancing generalization in natural language inference by syntax. In Findings of EMNLP.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

A. S. Hsu and N. Chater. 2010. The logical problem of language acquisition: a probabilistic perspective. Cogn. Sci., 34(6):972–1016.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1725–1744, Online. Association for Computational Linguistics.

James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1372–1379, Online. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in NLP: a taxonomy and review.

Dieuwke Hupkes, Sara Veldhoen, and Willem H. Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. Journal of Artificial Intelligence Research, 61:907—-926.

Najoung Kim and Paul Smolensky. 2021. Testing for grammatical category abstraction in neural language models. In Proceedings of The Society for Computation in Linguistics (SCiL).

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2962–2967, Copenhagen, Denmark. Association for Computational Linguistics.

Brenden M. Lake and Marco Baroni. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In 34th International Conference on Machine Learning.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Shalom Lappin and Stuart Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. Journal of Linguistics, 43:393–427.

Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. 2015. Unsupervised POS induction with word embeddings. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1311–1316, Denver, Colorado. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. Annual Review of Linguistics, 7:195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics, 4:521–535.

João Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 108–114.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? Experiments with Jabberwocky Probing. In NAACL-HLT.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. Computational Linguistics, 20(2):155–171.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Tiago Pimentel and Ryan Cotterell. 2021. A bayesian framework for information-theoretic probing. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4609–4622, Online. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Durme. 2018. Hypothesis only baselines in natural language inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 180–191.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3363–3377, Online. Association for Computational Linguistics.

Naomi Saphra and Adam Lopez. 2018. Language models learn POS first. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 328–330, Brussels, Belgium. Association for Computational Linguistics.

Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. 2022. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. Transactions of the Association for Computational Linguistics, 10:1423–1439.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In EMNLP-IJCNLP, pages 5830–5836. Association for Computational Linguistics.

Yikang Shen, Shawn Tan, Alessandro Sordoni, Siva Reddy, and Aaron C. Courville. 2020. Explicitly modeling syntax in language model improves generalization. ArXiv, abs/2011.07960.

Karl Stratos, Michael Collins, and Daniel Hsu. 2016. Unsupervised part-of-speech tagging with anchor hidden Markov models. Transactions of the Association for Computational Linguistics, 4:245–257.

Pang-ning Tan, Vipin Kumar, and Jaideep Srivastava. 2002. Selecting the right interestingness measure for association patterns. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the North American Chapter of the Association for Computational Linguistics, page 173–180.

Ke M. Tran, Yonatan Bisk, Ashish Vaswani, Daniel Marcu, and Kevin Knight. 2016. Unsupervised neural hidden Markov models. In Proceedings of the Workshop on Structured Prediction for NLP, pages 63–71, Austin, TX. Association for Computational Linguistics.

Josef Valvoda, Naomi Saphra, Jonathan Rawski, Adina Williams, and Ryan Cotterell. 2022. Benchmarking compositionality with formal languages. In Proceedings of the 29th International Conference on Computational Linguistics, pages 6007–6018, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.

Bowen Wu, Haoyang Huang, Zongsheng Wang, Qihang Feng, Jingsong Yu, and Baoxun Wang. 2019. Improving the robustness of deep reading comprehension models by leveraging syntax prior. In Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 53–57, Hong Kong, China. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints.

Yuan Yang and Steven T. Piantadosi. 2022. One model for the learning of language. Proceedings of the National Academy of Sciences, 119(5).

Appendix – Proofs and complements

Let p be a probability distribution, and f a mapping. One will denote $p \circ f^{-1}$ the probability distribution induced by f.

Remark 1. One has

$$\langle\!\langle \sigma(\boldsymbol{w}) \rangle\!\rangle = \langle\!\langle \boldsymbol{w} \rangle\!\rangle$$

$$\langle\!\langle f[\sigma(\boldsymbol{w}]) \rangle\!\rangle = \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle$$

$$f(\sigma(\boldsymbol{w})) = \sigma(f(\boldsymbol{w}))$$

$$p(f[\boldsymbol{w}]) = p \circ f^{-1}(f(\boldsymbol{w}))$$

$$p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) = p \circ f^{-1}(\langle\!\langle f(\boldsymbol{w}) \rangle\!\rangle)$$

Remark 2. One has

$$p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \frac{1}{\mu(\boldsymbol{w})} \sum_{\sigma \in \mathfrak{S}_n} p(\sigma(\boldsymbol{w}))$$

Remark 3. One has

$$p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) = \frac{1}{\mu(f(\boldsymbol{w}))} \sum_{\sigma \in \mathfrak{S}_n} p(f[\sigma(\boldsymbol{w})])$$

Proof. By Remark 2 applied to $p \circ f^{-1}$ and Remark 1. □

Remark 4. One has

$$\langle\!\langle \langle\!\langle \boldsymbol{w} \rangle\!\rangle \rangle\!\rangle = \langle\!\langle \boldsymbol{w} \rangle\!\rangle$$

$$f[f[\boldsymbol{w}]] = f[\boldsymbol{w}]$$

$$\langle\!\langle \sigma(\boldsymbol{w}) \rangle\!\rangle = \langle\!\langle \boldsymbol{w} \rangle\!\rangle$$

$$\langle\!\langle f[\boldsymbol{\sigma(w)}] \rangle\!\rangle = \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle$$

Proof. The second equality comes from

$$f^{-1} \circ f \circ f^{-1} \circ f = f^{-1} \circ f$$

□

Lemma 3. Let p be a distribution over A^+, and $f : A \mapsto B$ be a mapping. One supposes that there exists a mapping $\lambda_f(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$ such that $\forall \boldsymbol{w} \in A^+$

$$p(\boldsymbol{w} \mid \langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \lambda_f(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)\, p(f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)$$

then one has $\lambda_f(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \frac{\mu(\boldsymbol{w})}{\mu(f(\boldsymbol{w}))}$.

Proof. One has

$$\sum_{\sigma \in \mathfrak{S}_n} p(\sigma(\boldsymbol{w}) \mid \langle\!\langle \sigma(w) \rangle\!\rangle)$$

$$= \sum_{\sigma \in \mathfrak{S}_n} \lambda_f(\langle\!\langle \sigma(\boldsymbol{w}) \rangle\!\rangle)\, p(f[\boldsymbol{\sigma(w)}] \mid \langle\!\langle f[\boldsymbol{\sigma(w)}] \rangle\!\rangle)$$

$$= \frac{\lambda_f(\langle\!\langle (\boldsymbol{w}) \rangle\!\rangle)}{p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)} \sum_{\sigma \in \mathfrak{S}_n} p(f[\boldsymbol{\sigma(w)}])$$

$$= \lambda_f(\langle\!\langle (\boldsymbol{w}) \rangle\!\rangle) \mu(f(\boldsymbol{w}))$$

from Remark 1 and Remark 3. With the fact that

$$p(\sigma(\boldsymbol{w})) = p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) p(\sigma(\boldsymbol{w}) \mid \langle\!\langle \sigma(\boldsymbol{w}) \rangle\!\rangle)$$

one has

$$p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \frac{p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)}{\mu(\boldsymbol{w})} \sum_{\sigma \in \mathfrak{S}_n} p(\sigma(\boldsymbol{w}) \mid \langle\!\langle \sigma(w) \rangle\!\rangle)$$

$$= \frac{p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) \lambda_f(\langle\!\langle (\boldsymbol{w}) \rangle\!\rangle) \mu(f(\boldsymbol{w}))}{\mu(\boldsymbol{w})}$$

hence the result. □

Corollary 2. Let p be a distribution and f be a mapping. Then f factorises p iff

$$p(\boldsymbol{w}) = p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle (f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) \frac{\mu(\boldsymbol{w})}{\mu(f(\boldsymbol{w}))}$$

Lemma 4. Let p be a distribution and f a mapping. Then one has

$$\forall \boldsymbol{w} \in A^+, p(\boldsymbol{w} \mid f[\boldsymbol{w}]) = p(\boldsymbol{w} \mid A^{|\boldsymbol{w}|}) \Leftrightarrow$$

$$\forall \boldsymbol{w} \in A^+,\ p(f[\boldsymbol{w}]) = \begin{cases} p(A^{|w|}) & \text{or} \\ 0 \end{cases}$$

Proof. Suppose the left hand side of the equivalence, then, for any $\boldsymbol{w} \in A^+$, either $p(f[\boldsymbol{w}]) = 0$ or there exists $\boldsymbol{w'} \in f[\boldsymbol{w}]$ such that $p(\boldsymbol{w'}) > 0$ then $p(f[\boldsymbol{w}]) = p(f[\boldsymbol{w'}]) = p(A^{|w|})$.

On the other way, either $p(\boldsymbol{w}) = 0$ (and the equality holds) or $p(\boldsymbol{w}) > 0$ implying $p(f[\boldsymbol{w}]) > 0$ hence $p(f[\boldsymbol{w}]) = p(A^{|w|})$ implying in turn $p(\boldsymbol{w} \mid f[\boldsymbol{w}]) = p(\boldsymbol{w} \mid A^{|\boldsymbol{w}|})$. □

Computing Example 1. Let $\boldsymbol{w}_1$ be the sentence "cats eat rats" and let $\boldsymbol{w}_2$ be the sentence "men build houses". one has

$$\begin{array}{ll}
\langle\!\langle \boldsymbol{w}_1 \rangle\!\rangle = & \{\text{cats eat mice, cats mice eat,} \\
& \ldots, \text{mice eat cats}\} \\
\langle\!\langle \boldsymbol{w}_2 \rangle\!\rangle = & \{\text{men build houses,} \\
& \text{men houses build,} \\
& \ldots, \text{houses build men}\} \\
f(\boldsymbol{w}_1) = & b_0 b_1 b_2 \\
f(\boldsymbol{w}_2) = & b_0 b_1 b_2 \\
f[\boldsymbol{w}_1] = & \{\text{cats eat mice, cats build mice,} \\
& \ldots, \text{men eat mice,} \\
& \text{men build mice}\} \\
\langle\!\langle f[\boldsymbol{w}_1] \rangle\!\rangle = & \{\text{cats eat mice, cats build mice,} \\
& \ldots, \text{build mice men,} \\
& \text{men houses eat}\}
\end{array}$$

One has

$$p_{|f}(\boldsymbol{w}_1) =$$

$$p(\langle\!\langle \boldsymbol{w}_1 \rangle\!\rangle) p(f[\boldsymbol{w_1}] \mid \langle\!\langle (f[\boldsymbol{w_1}] \rangle\!\rangle) \frac{\mu(\boldsymbol{w}_1)}{\mu(f[\boldsymbol{w_1}])} =$$

$$\frac{1}{2} \cdot 1 \cdot \frac{1}{1}$$

hence f factorises p. One has

$$p(\boldsymbol{w}_1 \mid f[\boldsymbol{w_1}]) = \frac{\frac{1}{2}}{1} = p(\boldsymbol{w}_1 \mid A^3)$$

hence f is information-minimal for p.

The Lemma 5 gives a formula for the projection $p_{|f}$ of p conditionally to f.

Lemma 5. Let p be a probability distribution, and let f be a mapping. Let us define $\pi(p, f)$ by

$$\pi(p, f)(\boldsymbol{w}) = p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) p(f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) \frac{\mu(\boldsymbol{w})}{\mu(f(\boldsymbol{w}))}$$

then one has

$$p_{|f} = \pi(p, f)$$

One needs a few steps in order to prove Lemma 5.

Lemma 6. Let p be a distribution over A^+ and $f : A \mapsto B$ be a mapping. Then

1. $\pi(p, f)(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$
2. $\pi(p, f)(f[\boldsymbol{w}]) = p(f[\boldsymbol{w}])$

Proof. 1: one has

$$\pi(p, f)(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \frac{1}{\mu(\boldsymbol{w})} \sum_{\sigma \in \mathfrak{S}_n} \pi(p, f)(\sigma(\boldsymbol{w}))$$

with

$$\pi(p, f)(\sigma(\boldsymbol{w})) = \frac{p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) p(f[\boldsymbol{\sigma(w)}])}{p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)} \frac{\mu(\boldsymbol{w})}{\mu(f(\boldsymbol{w}))}$$

and

$$\sum_{\sigma \in \mathfrak{S}_n} \frac{p(f[\boldsymbol{\sigma(w)}])}{p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) \mu(f(\boldsymbol{w}))} = 1$$

hence the result.

2: one has

$$\pi(p, f)(f[\boldsymbol{w}])$$

$$= p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) p(f[f[\boldsymbol{w}]] \mid \langle\!\langle f[f[\boldsymbol{w}]] \rangle\!\rangle) \frac{\mu(f(\boldsymbol{w}))}{\mu(f(\boldsymbol{w}))}$$

$$= p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) p(f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) \frac{\mu(f(\boldsymbol{w}))}{\mu(f(\boldsymbol{w}))}$$

$$= p(f[\boldsymbol{w}])$$

□

Corollary 3. Let p be a distribution over A^+ and $f : A \mapsto B$ be a mapping. Then

1. $\pi(p, f)$ is a probability distribution.
2. $\pi(\pi(p, f), f) = \pi(p, f)$.

Proof. (1): The $\langle\!\langle \boldsymbol{w} \rangle\!\rangle$ form a partition of A^+ hence

$$\pi(p, f)(A^+) = \sum_{\langle\!\langle \boldsymbol{w} \rangle\!\rangle} \pi(p, f)(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$$

$$= \sum_{\langle\!\langle \boldsymbol{w} \rangle\!\rangle} p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = 1$$

(2): Since $\pi(p, f)$ is only computed from $p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$ and $p(f[\boldsymbol{w}])$, with the Lemma 6 one has the conclusion. □

Remark 5. In particular, one can give another definition of the set

$$\mathcal{F}_f = \{q \mid q \text{ is factorised by } f\}$$

as

$$\mathcal{F}_f = \{\pi(q, f) \mid q \text{ is a distribution}\}$$

Lemma 7. Let p and q be two probability distributions over A^+, and let $f : A \mapsto B$ be a mapping. Then one has:

$$H(p||\pi(q, f)) \geq H(p||\pi(p, f))$$

with equality iff $\pi(q, f) = \pi(p, f)$.

Proof. The inequality is equivalent to

$$\sum_{\boldsymbol{w}\in A^+} p(\boldsymbol{w}) \left(\log(\frac{q(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)}{p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)}) \right.$$

$$\left. + \log(\frac{q(f[\boldsymbol{w}])p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}{p(f[\boldsymbol{w}])q(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}) \right) \leq 0$$

By Jensen's inequality and concavity of the $\log$, and summing over $\langle\!\langle \boldsymbol{w} \rangle\!\rangle$ one has

$$\sum_{\boldsymbol{w}\in A^+} p(\boldsymbol{w}) \left(\log(\frac{q(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)}{p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)}) \right) =$$

$$\sum_{\langle\!\langle \boldsymbol{w} \rangle\!\rangle} p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) \left(\log(\frac{q(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)}{p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)}) \right) \leq 0$$

with equality iff $\forall \boldsymbol{w}, q(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$.

By summing over $\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle$, one has

$$\sum_{f[\boldsymbol{w}]} \frac{q(f[\boldsymbol{w}])p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}{q(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)} =$$

$$\sum_{\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle} \frac{q(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}{q(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)} = 1$$

and by Jensen's inequality and concavity of the $\log$, and summing over $f[\boldsymbol{w}]$ one has

$$\sum_{\boldsymbol{w}\in A^+} p(\boldsymbol{w}) \left(\log(\frac{q(f[\boldsymbol{w}])p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}{p(f[\boldsymbol{w}])q(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}) \right) =$$

$$\sum_{f[\boldsymbol{w}]} p(f[\boldsymbol{w}]) \left(\log(\frac{q(f[\boldsymbol{w}])p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}{p(f[\boldsymbol{w}])q(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)}) \right) \leq 0$$

with equality iff $\forall \boldsymbol{w}, q(f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle) = p(f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)$.

Because the value of $\pi(p, f)$ only depends on $p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle)$ and $p(f[\boldsymbol{w}] \mid \langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)$, the equality holds in the statement iff $\pi(q, f) = \pi(p, f)$. □

Proof of Lemma 5. One applies Remark 5 and Lemma 7, and one gets the result. □

Separating structure from data.

Our goal is to isolate structural information form contextual information for an observation $(a_1, \ldots, a_n)$.

For any permutation $\sigma \in \mathfrak{S}_n$, the tuple

$$(a_{\sigma(1)}, \ldots, a_{\sigma(n)}) = (a'_1, \ldots, a'_n)$$

satisfies a relation

$$(a_1, \ldots, a_n) = (a'_{\sigma^{-1}(1)}, \ldots, a'_{\sigma^{-1}(n)})$$

which will be denoted σ^{-1}.

Definition 7. For an observation $X = (a_1, \ldots, a_n)$, and a permutation $\sigma \in \mathfrak{S}_n$, let us denote

$$\sigma(X) = (a_{\sigma(1)}, \ldots, a_{\sigma(n)})$$

For any probability distribution p over A^+, one will define

$$p_2(X = (a_1, \ldots, a_n), Y = \sigma) = \frac{1}{n!}\, p(\sigma(X))$$

Lemma 8. One has

$$p(\boldsymbol{w}) = \sum_{\sigma\in\mathfrak{S}_n} p_2(X = \sigma^{-1}(\boldsymbol{w}), Y = \sigma)$$

$$p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \frac{|\boldsymbol{w}|!\; p_2(X = \boldsymbol{w})}{\mu(\boldsymbol{w})}$$

where $|\boldsymbol{w}|$ is the length of $\boldsymbol{w}$.

Proof. The first statement is just straightforward from the definition of p_2. In particular, one has

$$p(\boldsymbol{w}) = |\boldsymbol{w}|!\; p_2(Y = id, X = \boldsymbol{w})$$

One has

$$p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = \frac{1}{\mu(\boldsymbol{w})} \sum_{\sigma\in\mathfrak{S}_n} p(\sigma(\boldsymbol{w}))$$

$$= \frac{1}{\mu(\boldsymbol{w})} \sum_{\sigma\in\mathfrak{S}_n, \rho\in\mathfrak{S}_n} p_2(X = \rho^{-1}(\sigma(\boldsymbol{w})), Y = \rho)$$

$$= \frac{\sum_{\sigma\in\mathfrak{S}_n} p_2(X = \sigma(\boldsymbol{w}))}{\mu(\boldsymbol{w})}$$

and, with the fact that

$$\forall \sigma \in \mathfrak{S}_n, p_2(X = \boldsymbol{w}) = p_2(X = \sigma(\boldsymbol{w}))$$

one has the result. □

Definition 8. Let us define

$$p_{2|f}(X = \boldsymbol{w}, Y = \sigma) =$$

$$p_2(X = \boldsymbol{w}) p_2(Y = \sigma | f(X) = f(\boldsymbol{w}))$$

Lemma 9. One has

1. $p_{2|f} = p_{|f_2}$
2. $H(p_2 || p_{|f_2}) = H_{p_2}(Y|f(X)) + H_{p_2}(X)$
3. $H(p||p_{|f}) = H(p_2||p_{2|f}) - \boldsymbol{E}_p(\log(|\boldsymbol{w}|!))$

Proof. 1. One has

$$p_{|f_2}(X=\boldsymbol{w}, Y=\sigma) = \frac{1}{n!} p_{|f}(\sigma(\boldsymbol{w}))$$

$$= \frac{1}{n!} \frac{p_{|f}(\langle\!\langle \sigma(\boldsymbol{w}) \rangle\!\rangle) p_{|f}(f[\boldsymbol{\sigma}(\boldsymbol{w})])}{p_{|f}(\langle\!\langle f[\boldsymbol{\sigma}(\boldsymbol{w})] \rangle\!\rangle)} \frac{\mu(\sigma(\boldsymbol{w}))}{\mu(f(\boldsymbol{\sigma}(\boldsymbol{w})))}$$

with, by Lemma 6 and Lemma 8,

$$p_{|f}(\langle\!\langle \sigma(\boldsymbol{w}) \rangle\!\rangle) = p(\langle\!\langle \boldsymbol{w} \rangle\!\rangle) = n! \frac{p_2(X=\boldsymbol{w})}{\mu(\boldsymbol{w})}$$

and

$$p_{|f}(f[\boldsymbol{\sigma}(\boldsymbol{w})]) = p(f[\boldsymbol{\sigma}(\boldsymbol{w})])$$
$$= n! p_2(f(X) = f(\boldsymbol{w}), Y = \sigma)$$

and

$$p_{|f}(\langle\!\langle f[\boldsymbol{\sigma}(\boldsymbol{w})] \rangle\!\rangle) = p(\langle\!\langle f[\boldsymbol{w}] \rangle\!\rangle)$$
$$= n! \frac{p_2(f(X) = f(\boldsymbol{w}))}{\mu(f(\boldsymbol{w})}$$

and one has the result.

The second statement is an application of the definition of $p_{2|f}$ together with statement 1.

The third statement comes from the fact that

$$p(\boldsymbol{w}) = \sum_{\sigma \in \mathfrak{S}_n} p_2(X = \sigma^{-1}(\boldsymbol{w}), Y = \sigma)$$

$$p_{|f}(\boldsymbol{w}) = |\boldsymbol{w}|! p_{|f_2}(X = \sigma^{-1}(\boldsymbol{w}), Y = \sigma)$$

one has

$$H(p||p_{|f}) = -\sum_{\boldsymbol{w}} p(\boldsymbol{w}) \log(p(_{|f}(\boldsymbol{w}))$$

$$= -\sum_{\boldsymbol{w},\sigma} p_2(X = \sigma^{-1}(\boldsymbol{w}), Y = \sigma)$$

$$\log(|\boldsymbol{w}|! p_{|f_2}(X = \sigma^{-1}(\boldsymbol{w}), Y = \sigma))$$
$$= H(p_2||p_{2|f}) - \boldsymbol{E}_p(\log(|\boldsymbol{w}|!))$$

□

Lemma 10. Let p be a probability distribution, and let f and g be two mappings. One has

$$i_s(p, g \circ f) \leq i_s(p, f)$$

Proof. From Lemma 9, one has

$$i_s(p||f) = H_{p_2}(Y|z(X)) - H_{p_2}(Y|f(X))$$

and one has

$$H_{p_2}(Y|f(X)) \leq H_{p_2}(Y|g \circ f(X))$$

hence

$$i_s(p||f) \geq i_s(p||g \circ f)$$

□

Lemma 11. Let p be a probability distribution, and let f and g be two mappings. One has

$$i_s(p, g \circ f) \leq i_s(p, f)$$

Proof. With the fact that

$$H_p(f(\boldsymbol{w})) \geq H_p(g \circ f(\boldsymbol{w}))$$

one has the result. □

Details of Example 2 . The two sentences are strictly equivalent, thus we will only compute the values for, say, $\boldsymbol{u} =$ cats eat mice .

One has $p(\langle\!\langle \boldsymbol{u} \rangle\!\rangle) = p(\{cats, eat, mice\} = \frac{1}{2}$.

Let $z : A \mapsto \{a\}$ be a null mapping. By Lemma 5, one has

$$p_{|z}(\boldsymbol{u}) = p(\langle\!\langle \boldsymbol{u} \rangle\!\rangle) \frac{p(z[\boldsymbol{u}])}{p(\langle\!\langle z[\boldsymbol{u}] \rangle\!\rangle)} \frac{\mu(\boldsymbol{u})}{\mu(z(\boldsymbol{u}))}$$

with $p(z[\boldsymbol{u}]) = p(aaa) = 1$ and $p(\langle\!\langle z[\boldsymbol{u}] \rangle\!\rangle) = p(\{a, a, a\}) = 1$, $\mu(\boldsymbol{u}) = 1$ and $\mu(aaa) = 6$, one has $p_{|z}(\boldsymbol{u}) = \frac{1}{12}$ and finally

$$H(p \parallel p_{|z}) = \log(12)$$

One has also $p \circ z^{-1}(aaa) = 1$ thus

$$H(p \circ z^{-1}) = 0$$

One has $p_{|id}(\boldsymbol{u}) = \frac{1}{2}$ thus

$$H(p \parallel p_{|id}) = \log(2),\ i_s(p \parallel id) = \log(6)$$

and $p \circ id^{-1} = p$ thus

$$H(p \circ z^{-1}) = \log(2),\ i_c(p \parallel id) = \log(2)$$

Let g be the mapping

$$a = \{\text{cats, mice, men, houses}\}, b = \{\text{eat, build}\}$$

then one has

$$p_{|g}(\boldsymbol{u}) = p(\langle\!\langle \boldsymbol{u} \rangle\!\rangle) \frac{p(g[\boldsymbol{u}])}{p(\langle\!\langle g[\boldsymbol{u}] \rangle\!\rangle)} \frac{\mu(\boldsymbol{u})}{\mu(g(\boldsymbol{u}))}$$

with $p(g[\boldsymbol{u}]) = p(aba) = 1$, $p(\langle\!\langle g[\boldsymbol{u}] \rangle\!\rangle) = p(\{a, b, a\}) = 1$, $\mu(\boldsymbol{u}) = 1$ and $\mu(g(\boldsymbol{u})) = \mu(aba) = 2$ one has $p_{|g}(\boldsymbol{u}) = \frac{1}{4}$ and

$$H(p \parallel p_{|g}) = \log(4),\ i_s(p \parallel g) = \log(3)$$

One has $p \circ g^{-1}(aba) = 1$, thus

$$H(p \circ g^{-1}) = 0,\ i_c(p \parallel g) = 0$$

Let h be the mapping

$$c = \{\text{cats, eat, mice}\}, d = \{\text{men, build, houses}\}$$

then one has

$$p_{|h}(\boldsymbol{u}) = p(\langle\!\langle \boldsymbol{u} \rangle\!\rangle) \frac{p(h[\boldsymbol{u}])}{p(\langle\!\langle h[\boldsymbol{u}] \rangle\!\rangle)} \frac{\mu(\boldsymbol{u})}{\mu(h(\boldsymbol{u}))}$$

with $p(h[\boldsymbol{u}]) = p(ccc) = \frac{1}{2}$, $p(\langle\!\langle h[\boldsymbol{u}] \rangle\!\rangle) = p(\{c, c, c\}) = \frac{1}{2}$, $\mu(\boldsymbol{u}) = 1$ and $\mu(h(\boldsymbol{u})) = \mu(ccc) = 6$ one has $p_{|h}(\boldsymbol{u}) = \frac{1}{12}$ and

$$H(p \parallel p_{|h}) = \log(12),\ i_s(p \parallel h) = 0$$

One has $p \circ h^{-1}(ccc) = \frac{1}{2}$, thus

$$H(p \circ h^{-1}) = \log(2),\ i_c(p \parallel h) = \log(2)$$

One can check that the gain of information from z to g is purely structural, while the gain from z to h is purely contextual.

Proof of Proposition 1. From Lemma 9, one has

$$i_s(p||f) = H_{p_2}(Y|z(X)) - H_{p_2}(Y|f(X))$$

hence i_s is a sum of entropies, as well as i_c.

The Corollary 1 in (Antos and Kontoyiannis, 2001) states that for a countable support target distribution p with entropy H and its M.L.E. p_n with entropy $\hat{H}_n$, the plugin estimator satisfies

$$\lim_{n\to\infty} \hat{H}_n = H \quad a.s.$$

and this directly implies the conclusion. □

Lemma 12. Let p be a probability distribution, having a minimal factorization f. Then the intersection of all minimal syntaxes of p is a minimal syntax of p:

$$\mathcal{G}^*(p) = \cap_{f \in G^*(p)} \mathcal{G}(p, f) \in G^*(p)$$

Proof. Let f be an optimal factorizations of p. The property

$$\forall \boldsymbol{w} \in supp(p), p(\boldsymbol{w} \mid f[\boldsymbol{w}]) = p(\boldsymbol{w} \mid A^{|\boldsymbol{w}|})$$

means that there is only one observed pattern of categories of length $|\boldsymbol{w}|$, say $b_1 \dots b_n$.

For any $(i, n) \in \mathbb{N}^2$, let us define the subset

$$A_{(i,n)} = \{a \in A \mid \exists \boldsymbol{w} \in A^n, p(\boldsymbol{w}) > 0, w_i = a\}$$

Each subset $A_{(i,n)}$ is entirely inside a class of the partition induced by f – otherwise there would be at least two patterns for sequences of size n.

One builds a partition m of A by merging the subsets $A_{(i,n)}$ with non-empty intersection. Any class of m is entirely included in a class induced by f.

The mapping m is minimal by construction, and it is a refinement of f: $\mathcal{G}(p, f) \subseteq \mathcal{G}(p, f)$, and, by Lemma 10, m factorises p. □

Implicit disentanglement. We consider the following matrix of confusion

$$M = \begin{array}{c} \\ \overline{a} \\ a \end{array} \begin{array}{c} \begin{array}{cc} \overline{b} & b \end{array} \\ \begin{pmatrix} 0.202 & 0.049 \\ 0.212 & 0.537 \end{pmatrix} \end{array}$$

with the equivalence matrix (representing the fact that classifiers $(b) = (a)$)

$$M_{\Leftrightarrow} = \begin{array}{c} \\ \overline{a} \\ a \end{array} \begin{array}{c} \begin{array}{cc} \overline{a} & a \end{array} \\ \begin{pmatrix} 0.251 & 0 \\ 0 & 0.749 \end{pmatrix} \end{array}$$

and a matrix of an independent classifier (c) with probability μ of success is

$$M_{\perp\!\!\!\perp} = \begin{array}{c} \\ \overline{a} \\ a \end{array} \begin{array}{c} \begin{array}{cc} \overline{c} & c \end{array} \\ \begin{pmatrix} 0.251(1-\mu) & 0.251\mu \\ 0.749(1-\mu) & 0.749\mu \end{pmatrix} \end{array}$$

and we write M as a convex combination of $M_{\Leftrightarrow}$ and $M_{\perp\!\!\!\perp}$:

$$M = \lambda M_{\Leftrightarrow} + (1-\lambda) M_{\perp\!\!\!\perp}$$

This gives

$$\lambda = 0.522, \mu = 0.408$$

which leads to the interpretations:

- an independent process of POS identification involving HMM constraints has a 40% success rate.
- the rate of correct classification by (b) is 0.586, decomposed in 0.195 of success from an independent process, and 0.391 due to implicit disentanglement.
- 66.7% of the overall success rate can be imputed to implicit disentanglement.

The fraction of success due to implicit disentanglement is exactly the certainty factor $F = 0.667$

(see (Tan et al., 2002)) which represents a convex decomposition of M

$$M = F.M_{\Rightarrow} + (1 - F).M_{\perp\!\!\!\perp}$$

with:

$$M_{\Rightarrow} = \begin{pmatrix} 0.251 & 0.0 \\ 0.163 & 0.586 \end{pmatrix} M_{\perp\!\!\!\perp} = \begin{pmatrix} 0.104 & 0.147 \\ 0.310 & 0.439 \end{pmatrix}$$

where $M_{\Rightarrow}$ represents a complete implication and $M_{\perp\!\!\!\perp}$ represents a complete independence, both with same marginals as M. (see (Tan et al., 2002) for a definition in the context of association rules)