Introduction to NLP and Ambiguity

G. Poorna Prudhvi
Software Engineer
@poornaprudhvi

Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI).

The uniqueness in the language learning making it distinct from other forms of learning in artificial intelligence gave birth to specific field called natural language processing

The major concern is to make computer understand natural language and perform various useful tasks and also to give us a better understanding of language.

Humans and Natural Language

Human Cognition (Understanding) uses inbuilt socio cultural context and Knowledge about the world to learn the language.

Mother tongue is the language which we develop using our socio cultural context and with the context which we develop by this is utilised in learning other languages.

To make correct sense of what we are talking we make use of the knowledge about the world in phrasing the sentences

Knowledge of Language

Phonology – concerns how words are related to the sounds that realize them.

Morphology – concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language.

Syntax – concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.

Semantics – concerns what words mean and how these meaning combine in sentences to form sentence meaning. The study of context-independent meaning.

Pragmatics – concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

Discourse – concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, interpreting pronouns and interpreting the temporal aspects of the information.

World Knowledge – includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

Five general steps in nlp

Lexical Analysis – It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

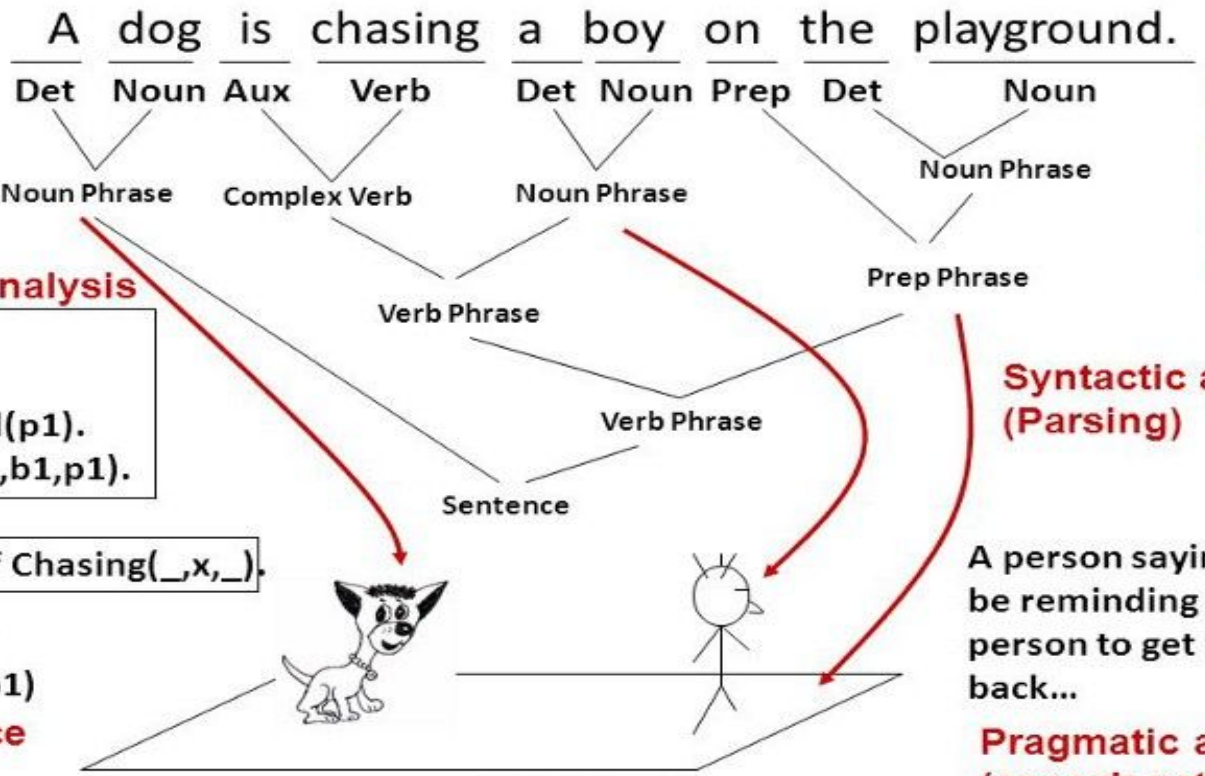
Syntactic Analysis (Parsing) – It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

Semantic Analysis – It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.

Discourse Integration – The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

Pragmatic Analysis – During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

An example of NLP



Components of NLP

Natural Language Understanding:

Mapping the given input in the natural language into a useful representation.
Different level of analysis required:
morphological analysis,
syntactic analysis,
semantic analysis,
discourse analysis, …

Natural Language Generation:

Producing output in the natural language from some internal representation.
Different level of synthesis required:
deep planning (what to say),
syntactic generation

Why Natural Language Understanding is hard?

- Natural language is extremely rich in form and structure, and very ambiguous.
  - How to represent meaning,
  - Which structures map to which meaning structures.
- One input can mean many different things. Ambiguity can be at different levels.
  - Meaning
  - Different ways to interpret sentence
  - Interpreting pronouns
  - Basing on context

What is ambiguity? why is it a problem?

- **Having more than one meaning**
- **Problems**
  - **combinatorial explosion**
    - Basically ambiguity increases the range of possible interpretations of natural language
    - Suppose each word in a 10 word sentence could have 3 interpretations. The number of interpretations of the whole sentence is going to be:
      - 3*3*3*3*3*3*3*3*3*3 = 59049

- Number of possible interpretations of a sentence will be more leading to difficulty in understanding the language

Local vs. global ambiguity

Global ambiguity:

The whole sentence can be interpreted in more than one form.

This needs semantic/pragmatic analysis to overcome

"I know more beautiful women than Kate"

(Let's guess 2 possible meanings here)

2 possible sentences are

"I know women more beautiful than Kate"

"I know more beautiful women than Kate does".

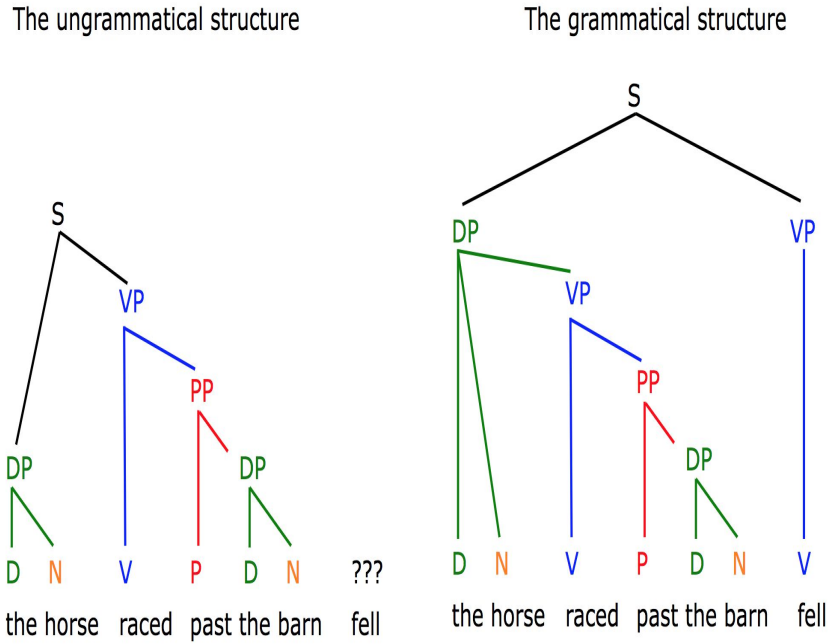
Local Ambiguity

The part of the sentences poses more than one interpretation in local ambiguity

Let's take look at this sentence

"The horse raced past the barn fell"

Can sometimes overcome by parsing



Funny Headlines

Sentence : The Pope’s baby step on gay

1. [The Pope's Baby]Step on Gays

2. [The Pope's] Baby Step on Gays

Lexical Ambiguity

Lexical ambiguity can occur when a word is polysemous, i.e. has more than one meaning, and the sentence in which it is contained can be interpreted differently depending on its correct sense.

Examples:

The word silver can be used as a noun, an adjective, or a verb.

She bagged two silver medals. [Noun]

She made a silver speech. [Adjective]

His worries had silvered his hair. [Verb]

How to resolve Lexical Ambiguity ?

Lexical ambiguity can be resolved by Lexical category disambiguation i.e, parts-of-speech tagging.

As many words may belong to more than one lexical category part-of-speech tagging is the process of assigning a part-of-speech or lexical category such as a noun, verb, pronoun, preposition, adverb, adjective etc. to each word in a sentence.

Parts of Speech Tagger

- A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word
- Classical POS tagging is done using CFG(Context Free Grammar) rules like
- Noun →flight | breeze | trip | morning | …
- Verb →is | prefer | like | need | want | fly
- Pronoun →me | I | you | it |
- Preposition →from | to | on | near | ..
- S →NP VP
- I want a morning flight

Lexical Semantic Ambiguity:

The type of lexical ambiguity, which occurs when a single word is associated with multiple senses.

Example: bank, pen, fast, bat, cricket etc.

The tank was full of water.

I saw a military tank.

The occurrence of tank in both sentences corresponds to the syntactic category noun, but their meanings are different.

Lexical Semantic ambiguity resolved using word sense disambiguation (WSD)

Word Sense Disambiguation

Computationally determining which sense of a word is activated by its use in a particular context.

E.g. I am going to withdraw money from the bank.

Knowledge Based Approaches

- Rely on knowledge resources like WordNet, Thesaurus etc.
- May use grammar rules for disambiguation.
- May use hand coded rules for disambiguation.

Machine Learning Based Approaches

- Rely on corpus evidence.
- Train a model using tagged or untagged corpus.
- Probabilistic/Statistical models.

Hybrid Approaches

- Use corpus evidence as well as semantic relations form WordNet.

WSD using parallel corpora

A word having multiple senses in one language will have distinct translations in another language, based on the context in which it is used.

The translations can thus be considered as contextual indicators of the sense of the word.

State of the art using word vectors

All the various sense of the words are represented as different vector in the vector space and the ambiguity is resolved.

Word Vectors

Representing word as a vector of numbers

A vector can be represented in a way we want like the values in vector may correspond to the document the word contains, the position of word, neighbouring words.

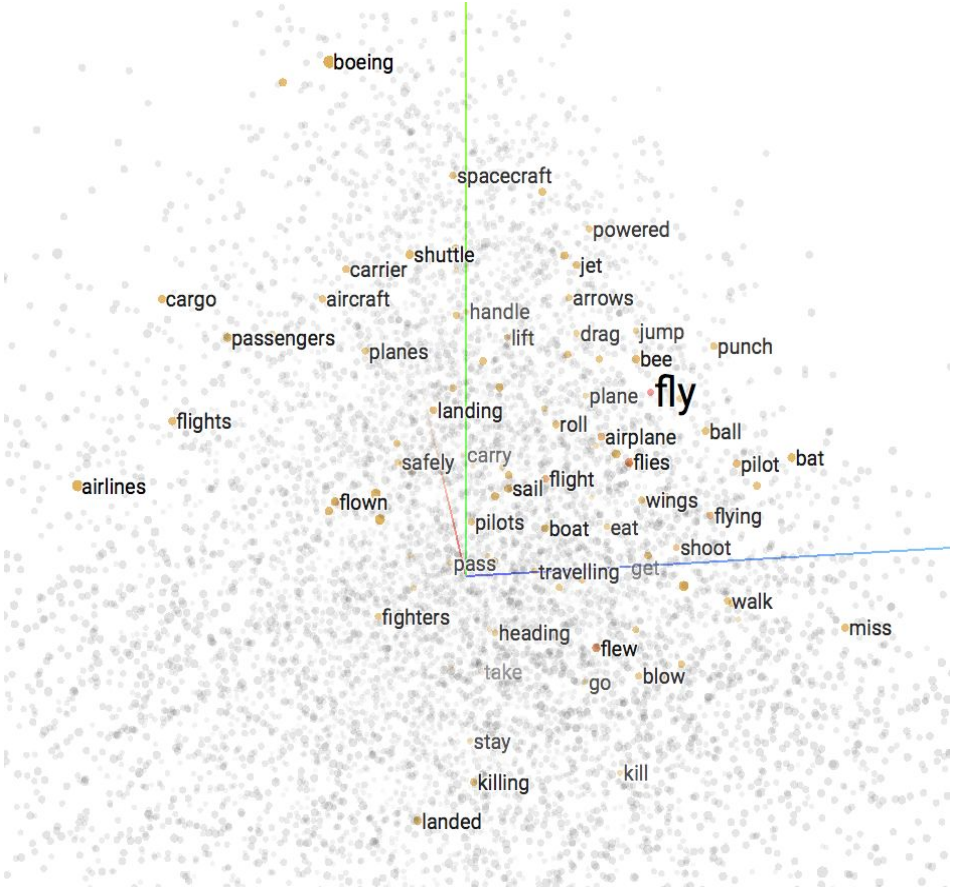
These vectors are of high dimension and are sparse. The dense representation of words are called embeddings.

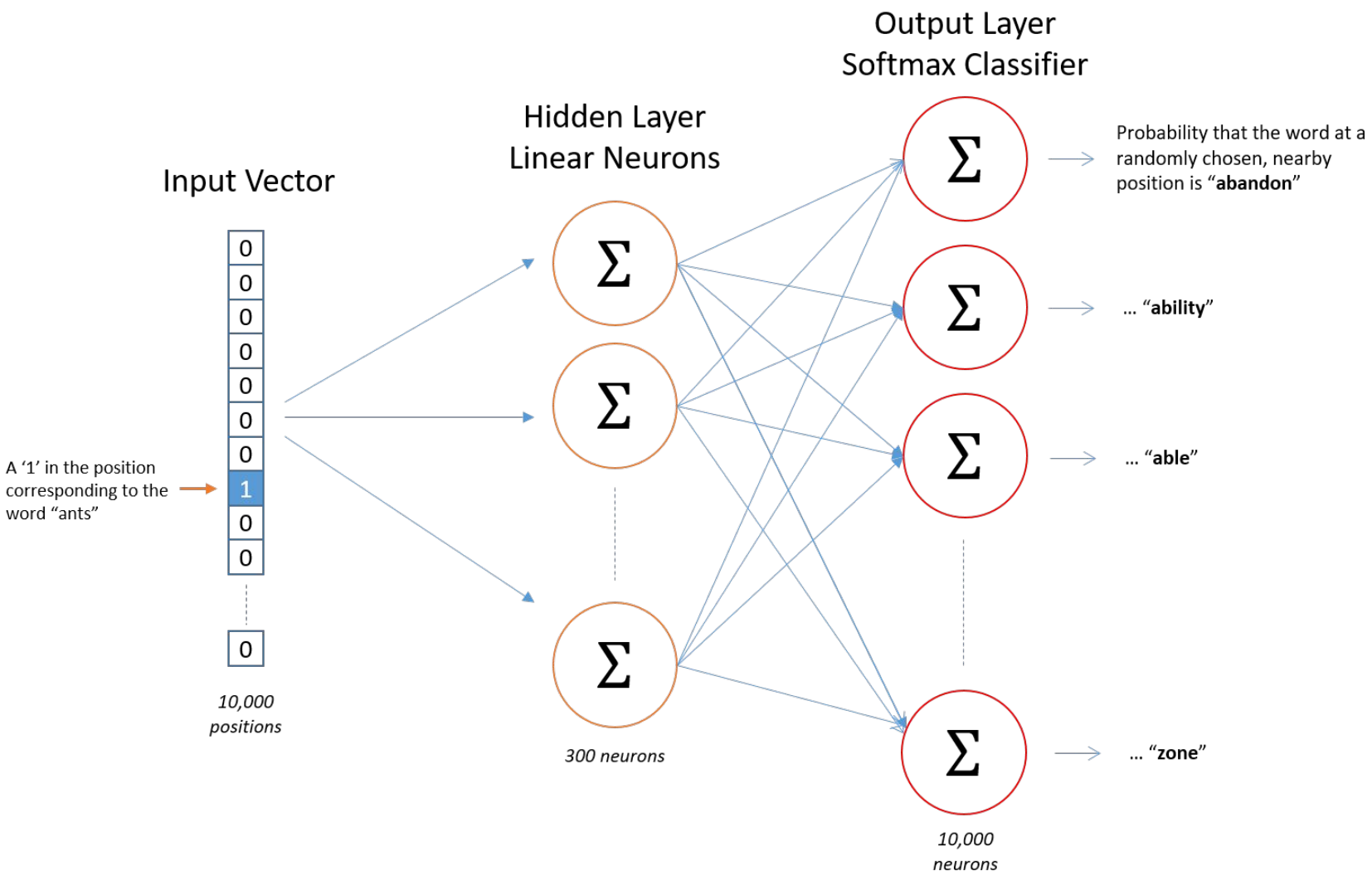
These embeddings can be learnt by Matrix Factorization or Neural Networks.

You can see the words distributed over a high dimensional space like this.

The words which are in same context will be nearer

Example : spacecraft, shuttle, jet are nearer to each other




Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
10,000 neurons

Acronyms

While many words in English are polysemous, things turn absolutely chaotic with acronyms. Acronyms are highly polysemous, some having dozens of different expansions.acronyms are often domain-specific and not commonly known.

Example: NLP → Neuro Linguistic Programming or Natural Language Processing

Given enough context (e.g. "2017" is a context word for the acronym ACL), it is possible to find texts that contain the expansion. This can either be by searching for a pattern (e.g. "Association for Computational Linguistics (ACL)") or considering all the word sequences that start with these initials, and deciding on the correct one using rules or a machine-learning based solution.

Syntactic Ambiguity

This refers to ambiguity in sentence structure and be able to interpret in different forms.

Examples:

They ate pizza with anchovies

I shot an elephant wearing my pajamas

Common to all these examples is that each can be interpreted as multiple different meanings, where the different meanings differ in the underlying syntax of the sentence.

The first sentence

"They ate pizza with anchovies", can be interpreted as

(i) "they ate pizza and the pizza had anchovies on it",

(ii) they ate pizza using anchovies

(iii) they ate pizza and their anchovy friends ate pizza with them.

The second sentence

"I shot an elephant wearing my pajamas" has two ambiguities:

first, does shoot mean taking a photo of, or pointing a gun to? (a lexical ambiguity).

But more importantly, who's wearing the pajamas?

Is it the person or the elephant

Existing Solutions for Syntactic Ambiguity

In the past, parsers were based on deterministic grammar rules (e.g. a noun and a modifier create a noun-phrase) rather than on machine learning.

Now parsers are mostly based on neural networks. In addition to other information, the word embeddings of the words in the sentence are used for deciding on the correct output. So potentially, such a parser may learn that "eat * with [y]" yields the output in the left of the image if y is edible (similar to word embeddings of other edible things), otherwise the right one.

Referential Ambiguity

Very often a text mentions an entity (someone/something), and then refers to it again, possibly in a different sentence, using another word.

Pronoun causing ambiguity when it is not clear which noun it is referring to.

Examples:

John met Mary and Tom. They went to restaurant [is they referring to mary and tom or all of them?]

John met Bill before he went to store [is he john or bill?]

Existing Solutions for Referential Ambiguity

Coreference resolution used to overcome referential ambiguity

A typical coreference resolution algorithm goes like this:

- extract a series of mentions [words referring to entities]
- For each mentions and each pair of mentions, compute a set of features
- Then, we find the most likely antecedent for each mention (if there is one) based on this set of features.

Sentence: My sister has a dog and she loves him very much

1. My sister has a dog and she loves him very much
2. Sister → Female, Dog → animal, she → feminine pronoun, him → masculine pronoun
3. My sister has a dog and she loves him very much

State of the art coreference resolution parser is by spacy.io using deep learning

Ellipsis

Incomplete sentence where missing item is not clear

Example:

"Peter worked hard and passed the exam. Kevin too"

Three possible interpretations of this example are

- Kevin worked hard
- Kevin passed the exam
- Kevin did both

Syntactic and semantic analysis might help.

Noun Compounds

English allows long series of nouns to be strung together using the incredibly ambiguous rule NG -> NG NG.

E.g. "New York University Martin Luther King Jr. scholarship program projects coordinator Susan Reid". Even taking "New York" "Martin Luther King Jr." and "Susan Reid" to be effectively single elements, this is 8 elements in a row, and has 429 possible parses.

Existing Solutions for Noun-compound Interpretation

(1) machine-learning methods like hand-labeling a bunch of noun-compounds to a set of pre-defined relations (e.g. part of, made of, means, purpose...), and learning to predict the relation for unseen noun-compounds.

(2) define a set of semantic relations which capture the interaction between the modifier and the head noun, and then attempt to assign one of these semantic relations to each compound.

For example, the compound phrase flu virus could be assigned the semantic relation causal (the virus causes the flu); the relation for desert wind could be location (the wind is located in the desert).

(3) Using Knowledge base and classifying differential compound nouns

Let’s try this now.

“I made her duck.”

How many different interpretations does this sentence have?

Solution

Some interpretations of : I made her duck.

I cooked duck for her.

I cooked duck belonging to her.

I created a toy duck which she owns.

I caused her to quickly lower her head or body.

I used magic and turned her into a duck.

With all these in mind….

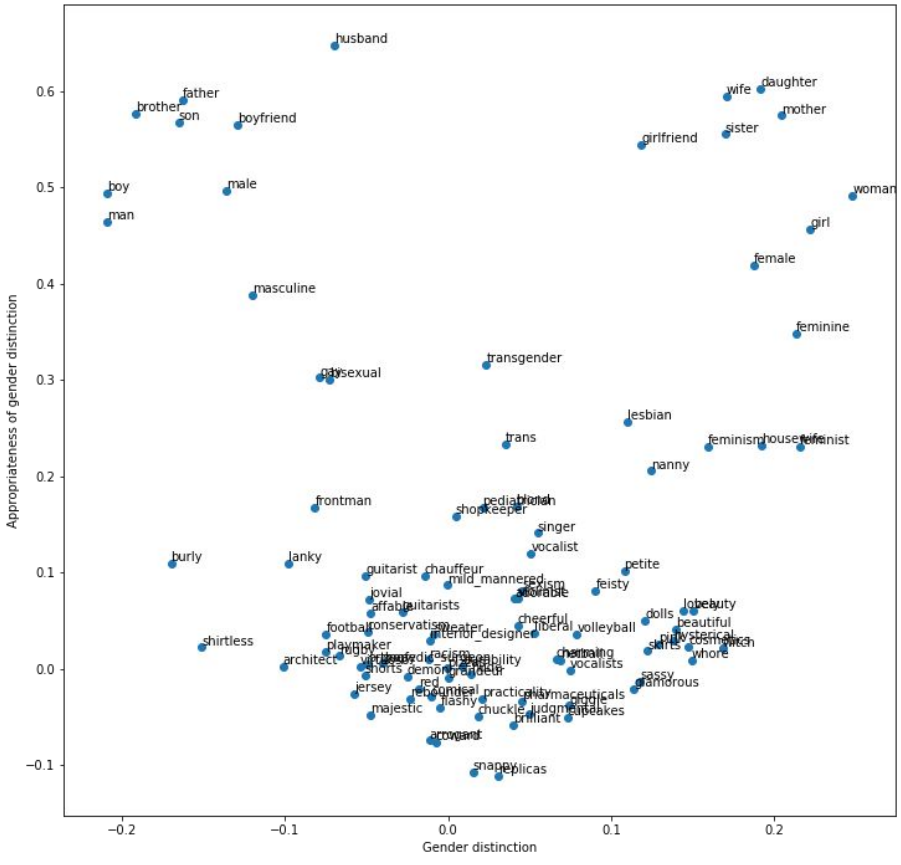
I would like to tell on more problem we are getting through context that is bias.

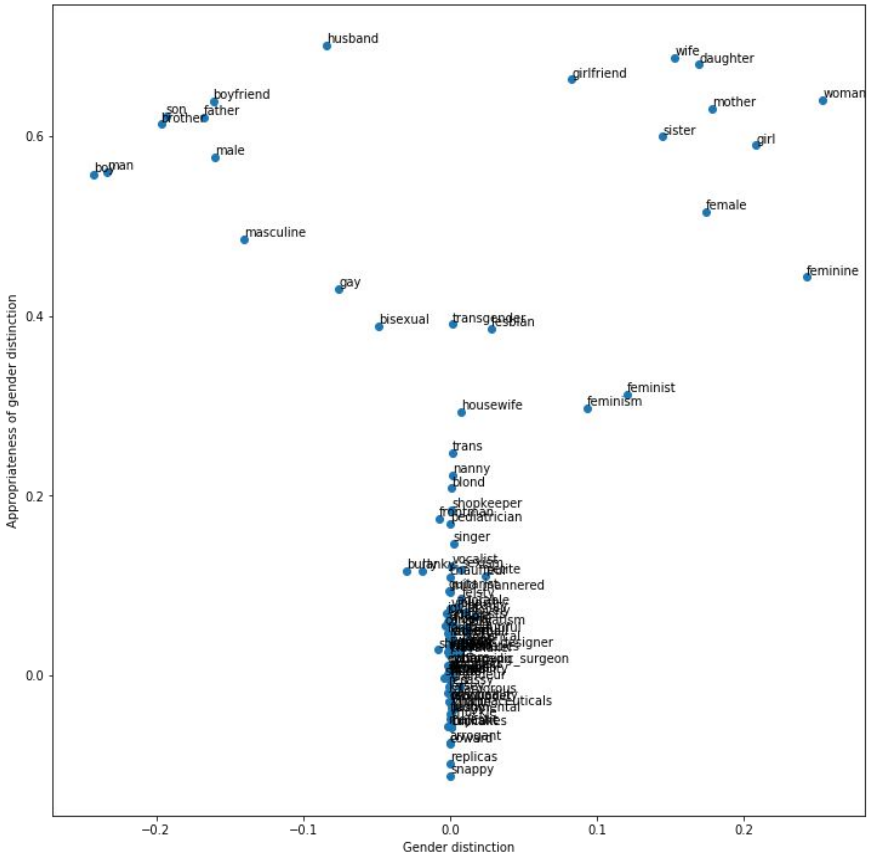
The data on which we are training the models are biased so is the models. There should be some mechanism to debias the models.

Example:

Word2vec biased because of its training on news data

Conceptnet debiased using some debiasing techniques


husband
father
brother
son
boyfriend
wife
daughter
mother
sister
girlfriend
boy
male
woman
man
girl
female
masculine
feminine
transgender
gay
bisexual
lesbian
trans
feminism
housewife
feminist
nanny
frontman
shopkeeper
singer
vocalist
burly
lanky
guitarist
chauffeur
petite
mild_mannered
jovial
feisty
beauty
guitarists
cheerful
dolls
beautiful
conservatism
theater
liberal
volleyball
shirtless
interior_designer
playmaker
racism
whore
architect
vocalists
sassy
glamorous
jersey
practicality
majestic
chuckle
brilliant
arrogant
snappy
replicas
Appropriateness of gender distinction
Gender distinction


husband
wife
daughter
girlfriend
boyfriend
mother
woman
son
father
brother
sister
girl
male
boy
man
female
masculine
feminine
gay
bisexual
transgender
lesbian
feminist
feminism
housewife
trans
nanny
blond
shopkeeper
pediatrician
singer
petite
mild_mannered
jovial
feisty
beauty
beautiful
hysterical
dolls
interior_designer
whore
vocalists
sassy
glamorous
practicality
majestic
chuckle
brilliant
arrogant
snappy
replicas
Appropriateness of gender distinction
Gender distinction

Thank you :)