What is Natural language processing?

It is a field that belongs to computer science, artificial intelligence and computational linguistics. The focus is the interaction between a computer and the human language and how the computer can understand a given text. It does so by parsing the input given and applying algorithms to figure out what the words mean and how they are related to each other.

The language concept that the computer must take care of can be sorted into speaking, reading and writing. By using that language provided the computer can understand how humans take decisions, plan their actions as well as even dreaming. Thus, understanding how we process and understand language is a key how the brain works. Recent NLP is heavily based on neural networks and deep learning.

Recent development in NLP is heavily focused on vector spaces. All words of a given data sets are stored and live inside a semantic multidimensional space which represents memory in our brain (according to theory of course). Words are presented as vectors that have attributes with values based on similar or related words, which means that that words can have a vector distance to other words. Those who are related live closer together in that space and those who aren't live further away.

Here is an example. A woman is to man as queen is to a king. The queen word is a vector built by attributes that form a concept. So a concept is made out of different words based on other words like features. A queen can be made out of different words or attributes that are related to that particular concept. Queen might have the attributes woman=1, royal = 0.95, rich=0.8, power=0.7 and so on. So by taking the

meaning(queen) – meaning(woman) + meaning(man)

Can be translated to

(woman=1 + royal = 0.95 + rich=0.8 + power=0.7) – (woman=1) + (man=1) which would give us

man=1 + royal = 0.95 + rich=0.8 + power=0.7 = meaning(King)

Thus the vector space can define a distance between words based on the values of the attributes. The same method can be applied to other concepts, just like Stockholm – Sweden + Iceland would give us Reykjavík. From a Cambridge lecture (linked below) the crowd was asked about what words they would use in similar situations. If they had the word Einstein and his profession is scientist, the question was what Messi's profession is. The crowd replied footballer but computer said midfielder. More examples are presented in the following table

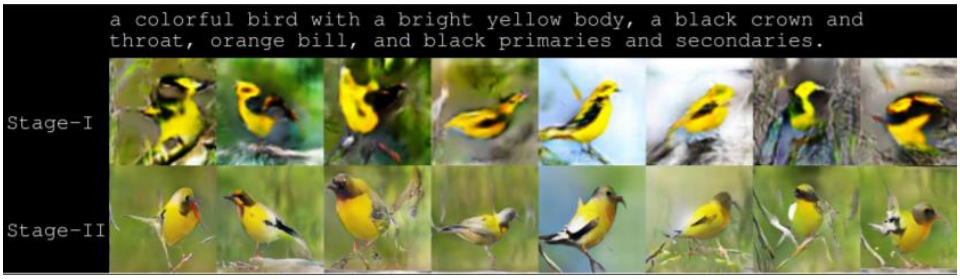
Word given	*Guessed by crowd*	*Guessed by computer*
Japan -> Sushi	-	-
German -> ???	Bratwurst	Bratwurst
France -> ???	Baguette	Tapas
USA -> ???	Burgers	Pizza

So it seems that this method doesn't follow human reasoning but rather logic, which is quite close to us humans. In the France case, people agreed that baguette was most appropriate because that is the picture most people have of it but the computer might base it on most eaten food in that country. We learn words from experience, so must the computer learn from data.

So with this method the computers can write text by prediction based on sequences of older words or sentences. They learn how words are related together or have similar synonyms. AI has been used to write news about certain topic, provided by facts and statistics.

They have been used in computer vision as well to find pictures based on vector combinations. I'm just going to post this here since a picture says more than a thousand words.

A really interesting approach is using GAN networks to creating images based on text.



The last interesting library that I have seen is Word2vec, which greatly helps putting words and their meaning to vector spaces. I recommend watching this lecture.

Question 1

Explain why a vector based NLP says that a wolf + house pet might suggest that the most appropriate word is dog, rather than a cat?

The wolf and a dog might share similar attributes in their vector space like having four legs, furry, bark, carnivores, live in hierarchy etc. By adding pet attributes to a wolf like friendly, loving, part of human family or more the distance of the word is most likely closer to a dog rather than a cat. By changing the word wolf to a tiger, a cat might probably be more related.

Question 2

Discuss these two following sentences and argue whether they are generated by a computer or a human.

"More than 1 million people bought the new iPhone this weekend"

I'd say this is most likely a computer since it only provides us with solid facts.

"The new coffe at starbucks surely warms your body and feelings like the hug of a nice lover"

For this I'd say that a human wrote this, since metaphors are used, words are spelled incorrectly, Starbucks is written with a lower s and what is a nice lover?

Note: I wrote both of them but tried to make the former one as *dry* as I could. This is really debatable.

Question 3

Name at least three different types of real world applications where NLP is used.

They are currently being used in Gmail to automatically create short responses to emails such as confirmation to an event or approving some question. They are being used in chatbots to find out what the users are saying and replying with the most appropriate answer. Finally they are used in Google search so when a user is typing in something, it suggests the next word for the user based on former input.

Question 4

NLP has been used to fight off spam emails and text messages. In what way can the NLP find out of the message is a spam or not?

By viewing all the words in an email, it could compare that values of those words and their concept to a trained set of correlated words in other known spam emails. For an example, if a message contained the sequence "WIN FREE CAR", that spam prediction score might be higher than the text from a friend "Hey man, I'm free next Saturday. Want to go and help me buy a car? I'll buy you some pizza. Win win for both of us". There is less distance between the classified spam words in first sentence versus the latter.

Question 5

Argue if a NLP could be used to identify or create new sets of music, if it would be trained on a specific type of genre.

A music is surely a form of language so it could definitely be used for identification and generation. However, the NLP should be trained on similar artists that follow similar music patterns or roles. Artists and their music can surely differ with styles, tempos and length of tracks as well as if songs are similar in structure. Some songs come in similar forms (intro, chorus, outro etc) while some are more static and same (hip hop has the same beat vs symphony is quite diverse).