Олимпиада для студентов и выпускников – 2018 г.

Направление: «Фундаментальная и прикладная лингвистика»

Профиль: «Компьютерная лингвистика» КОД - 310

Время выполнения задания – 180 мин., язык – русский.

Задание 1.

В социальной сети «Свиттер» каждый пользователь может подписаться на обновления любого другого пользователя. Всего в «Свиттере» зарегистрировано 2018 человек. Исследователь социальных сетей Т. А. Рантул знает, что в «Свиттере» есть пользователь, который ни на кого не подписан, но на него подписаны все остальные, и хочет его найти. Для этого он написал программу, которая умеет делать запрос к «Свиттеру» вида «подписан ли пользователь X на обновления пользователя Y?». Какое минимальное число запросов такого вида нужно будет сделать программе, чтобы наверняка найти того, кого она ищет? а) Докажите, что этого числа хватит; б) Докажите, что любого меньшего может не хватить.

Задание 2.

Перед вами стоит задача разработать диалоговую систему. Система должна распознавать “дубликаты” вопросов (дубликаты и нечеткие дубликаты), т.е. вопросы, на которые можно иметь одинаковые ответы. В частности, в систему должен войти некоторый модуль, который по предъявленным двум вопросам определяет, являются ли вопросы дубликатами (похожими вопросами).

Например, одинаковыми считаются вопросы «Где можно найти хорошие насосы?» и “Место продажи насоса” или “Как узнать, как решить задачу на сингулярное разложение матрицы?” и “Где найти пример решения задачи на SVD”.

Задача: для приведенных ниже пар вопросов определите, являются ли они перифразами. Опишите методы АОТ, с помощью которых можно было бы определить, что пара является перифразами.

Для этого:

(а) разметьте пары вопросов: ок – перифраза, нет – не перифраза.

Для каждой из пар, которые Вы считаете перифразами укажите:

(б) укажите какую задачу обработки текста необходимо, чтобы программа решала, чтобы можно было вычислить близость соответствующих вопросов (например, для пары из примера достаточно стемминга глагола, снятия морфологической омонимии, определения падежей …); это может быть спелл-чекинг, обработка буквенно-цифровых комплексов, распознавание подлежащего в предложении; выделение коллокаций с использованием взаимной информации, векторизация с использованием idf….

(в) приведите пример 2-3 правил, которые должны содержаться в модуле распознавания дубликатов, если этот модуль основан на правилах (правило не обязательно должно быть применимо ко всем данным правильным результатом)

(г) приведите пример признаков, которые следует учесть, если в модуле применяется машинное обучение; что должно являться объектами в системе машинного обучения, что – признаками, какой метод стоит применить (например, классификатор Байеса, деревья решений, кластеризацию ….)

(д) предложите 3 своих пар вопросов, которые можно дубликатами, которые «ловятся» методами, не упомянутыми в комментариях к вопросам из задания; укажите метод

NB!!! При формулировке Ваших правил, признаков, описаний типов обработки не забывайте, что не все пары являются нечеткими дубликатами;

Вопросы для распознавания:

	вопрос 1	вопрос 2	Перифраза/нет
1	Не открывается csv файл	чем открыть .csv файл?	
2	что подарить отцу	как выбрать подарок для папы?	
3	Какой RGB код у светло-голубого?	Какой RGB код у светло-зелёного?	
4	Где можно покататься с другом на лыжах и вкусно поесть?	В каком месте можно покататься с другом на лыжах или вкусно поесть.	
5	Почему ракеты красят в белый?	Почему ракеты всегда белые?	
6	Как написать чат-бота на python?	Чат бот на питоне	
7	Открыт ли сегодня исторический музей	Работает ли исторический музей завтра?	
8	Как подать заявлене о пропаже собаки?	Куда и кому заявить о том, что пропал пес?	
9	О чём замечательная книга Дж.Роулинг?	Содержание прекрасной пьесы Роулинг?	
10	Где находится ресторан Армения	Ресторан Հայաստան расположение на карте	
11	как упростить изучение статистики?	как сделать обучение статистике проще?	
12	Как упаковать подарок?	Как распаковать подарок?	
13	Кто из студентов 4-ого курса находится на стажировке и в экспедиции?	Кто из студентов четвертого курса находится на стажировке или в экспедиции?	
14	Как узнать, это лист клена, дуба или чего-то еще?	Как определить листья разных деревьев?	

Задание 3.

Прочтите пост из блога, посвященного автоматической обработке естественного языка (текст приведен ниже). Составьте краткое резюме этого поста на русском языке (объемом в один абзац, но не более 1000 знаков), отразив в нем основной тезис автора поста. Выскажите ваши собственные соображения по проблемам, затронутым в посте и их решениям.

On-Device Machine Intelligence

One problem with automated sentiment analysis is the limit of real-world human-based consensus on sentiment.

I think the recent thought is that humans usually agree about 80% of the time... which means software can only approach that. If we restrict the domain (which makes inferring context easier and allows us to build more biased models; i.e., we can assume more consensus between competing judgements) then you can do better.

I think more work in social science needs to be done, or NLP needs to dig more into the literature, on how humans agree on various topics within certain domains and relative to the context of those people making judgements (even movie reviews are crazy hard! take a simple binary "Bad|Good" boundary... and depending on the people you ask you get crazy different overall groupings about whether X movie was good or bad!!).

Take polarizing topics in specific context and you can get obvious boundaries (like abortion at a GOP conference, etc...). Widen the context (i.e., number of competing judgements from a more diverse sample size... so now mix people from other political parties in other countries), and you get messier boundaries.

In short, I think, sentiment analysis is not constrained by the technology, but by human behavior in general. Sentiment analysis works well if applied under the right conditions.... The question is, are those conditions so constrained by bias that it's even worth the effort using such a technique.

tldr;

Basically, you would need a model of human judgments per topic, which requires a model of what humans use to make judgements. And what do we use? In reality, it is a very sticky domain. Not even the best social science does this well (recent studies in social science are finding that long-standing "facts" about humans are clearly biased as they were based on research done against people coming from very similar cultures and backgrounds.... you need good sample size with the right distribution if you wanna study things like "What is humor?".... and the funding for such massive and longitudinal studies in social science is not there... it may never).

Not only that, say you have a highly constrained domain ("What is the judgement from other football fans if I say this about topic X"), you need a good model for how those specific judgements are built up. What if topic "X" is domestic abuse? What if it is about "deflategate" (which is really about a form of cheating)... these are not simple things. Even the performance of an athlete in one game is mired in socio-economic and racial contexts (sure if it's just statistics and things like yards per carry... but we are talking about human judgements... and like it or not humans use some pretty messy stuff to make judgements about other humans... including simple texts taken out of context). And, personally, we are a long way from understanding anything significant about human behavior, relations, and judgements. So you can't just tease out "sentiment" as a pattern in the textual data... not in the same way you can with things like word order, word counts, etc....

Even simple things like "Do people find X funny?" Who are the people? What was the topic? Just look a the difference in race and comedy, both the race of comics and their main audience. Or do the same thing with culture. In fact, humor is not simple at all.... though after the fact it's pretty easy to see that some people thought X was funny and others did not. This gives the impression that it easy to model. All sentiment analysis works like this, in my opinion. We see judgements after the fact, and see a pattern, and so we assume the pattern is predictive.

tldr;

For example, using SA for sports analysis: in the context of 2 competing teams, applied to the texts of those fans just from the two competing teams it's kind of a given where the boundary is. All you get from the Sentiment Analysis is a fairly quick easy way to sort messages into groups.

But do it with something like American football Superbowl: a large part of the text is coming from people who are not distributed into 2 simple groups (for or against team A) based on a previous bias. So you need to tease out the grouping. But the boundaries between competing judgements gets messier PLUS you need to start taking into account a much large domain of context. So large, in fact, that you may get poor model. Not only that, many people who are not

really football fans watch the superbowl and make judgements (say you are scraping twitter comments). Also, the intuition here, and the bias in the model (without more evidence to show), is that you are looking, during the regular football season, at a very specific Male population that probably reflects a fairly accurate distribution of race. Who knows about socio-economics (do rich white males what football??, what about poor black males??).... THE POINT IS, IF YOU WANT AN ACCURATE MODEL YOU CAN'T JUST ASSUME THE GENDER OR RACE OR SOCIO-ECONOMICS OF THE SPEAKERS... you need a good model of that too.

Some recent work on things like determining gender from food reviews is, I think, a better way to approach it. Instead of targeting sentiment as the goal, instead try to tease out certain attributes, and use those attributes to build up more abstract models where you can kinda guess the sentiment. This has the benefit of building up some context based on the attributes you model. Plus, it's more like what many humans do.... before making a judgement on speaker Xs' sentiment we do filter that through a lot of bias (for better or worse)... including personal relationship, gender, race, age.... Although maybe not politically correct, we do it nonetheless.... which makes building statistical models in this regard tricky.

For example, and again, good social science should be used here, but take many divisive topics on race and gender in the real world. We expect the dominant group to display much different sentiment about real world events (like the question "Is (racism/sexism/ageism) still a problem"...) distinct from the affected group. To get a good real-world model you need to account for this. But how do you build a model that is inherently 'racist/sexist/ageist' without causing a lot of problems, namely defending the accuracy of the model?

Other less divisive domains have the same issues. Socio-economic issues, or things like sports or fashion (where you expect not to have to deal with divisive things like race or religion) still have these problems. Consensus across diverse sample groups will not be very high, and relative to the theme/topic of the, may vary greatly.

In short, again, I don't think there is any technical limitation in sentiment analysis... pretty much the same techniques applied to other areas of NLP work in sentiment analysis. The issue is more a limit on human consensus than building models to good accuracy. So sentiment analysis works pretty good in very small domains.

https://www.quora.com/How-could-AI-NLP-sentiment-analysis-be-used-to-predict-whether-or-not-people-will-judge-you-based-on-something-you-say/answer/Joshua-Bowles.