**Sentence Topics**

*This paper deals with a preliminary version of the new sentence extraction algorithm which, rather than relying on just the presence of a single word or a set of words in the same sentence, firstly extracts the main sentence topics, than checks if particular keywords are present in the sentence extracted topics.* *The aim of the whole procedure is to avoid picking up sentences in which the keyword is just present without being the core subject. This part of the research focuses on sentence topics, but some work has already been spent on sentences extraction.*

*Introduction*

Consider the following sentence:

*“Wheat market is showing a positive trend: trained by last year good American production, wheat price arose more than prices of rice and soybeans.”*

This sentence shows the words wheat, soybeans and rice. By logic however this sentence deals with the state of wheat price, therefore it’s linked more to the wheat commodity, rather than rice and soybeans which are present just in comparative terms.

Using the previous version of the sentence extractor the following sentences would have been extracted

|  |  |
| --- | --- |
| **Using the set of keywords:** | **Extracted Sentence** |
| Wheat Keywords | *“Wheat market is showing a positive trend”, “wheat price arose more than prices of rice and soybeans”* |
| Rice Keywords | *“wheat price arose more than prices of rice and soybeans”* |
| Soybean Keywords | *“wheat price arose more than prices fo rice and soybeans”* |

The phallacy is given by the fact that the word *“rice”* as the word *“soybeans”* are present in a sentence that mostly deals with wheat.

In order to solve the issue the main topics of the sentence are extracted using a new algorithm. Sentence topic-keywords are:

*'Wheat', 'positive trend', 'year good', 'American production', 'wheat price arose'*

By looking at sentence topic-keywords instead of words, the fact that the sentence is mostly related to wheat appears clear: rice and soybeans do not even appear.

The main proposal is to look at sentence topics instead of single words for the relevant to one entity sentences selection process.

*The Method*

Sentence Topics extraction relies on Part Of Speech tagging, namely the recognition of each word role in the sentence: words will be labeled as nouns, verbs *et cetera* depending on their logical function. Obviously the algorithm is not that straightforward, because it relies on a tagged corpora, from which are then trained other taggers.

NLTK comes with a few tagged corpora, namely a dictionary of words and respective logical function labels. By the way tagged words cannot be used into sentence topics extraction because, by using just a dictionary of tagged words, each word logical function will be insensitive of the word role in the sentence.

In order to avoid this the algorithm works using the *tagged sentences* corpus from the Brown corpus.

First of all a Regular Expression Tagger is built using a list of words of suffixes which give to each word its function label, namely common noun, verb, adjective et cetera. The RE tagger first tries on the word all the possible labels then, as a residual option if no labels fit, at last assigns the label “Noun”.

Then a Unigram tagger is trained on the same corpus using the Regular Expression Tagger as backoff. A Unigram tagger is a tagger that still doesn’t evaluate the role of the word in the sentence, but just gives to the word the most probable label where probabilities are defined using the corpus.

As last preliminary step a Bigram tagger is trained using the same Brown corpus and the Unigram tagger as backoff. The Bigram tagger considers words in pairs when assigns the labels, therefore they are given depending on the (small) context defined by the words pair.

Once the words are labeled depending on their logical function, A Context Free Grammar is defined: in the CFG a few of the possible interactions between the words are modeled. In particular the interactions involving singular and plural common nouns, proper nouns and the interaction between adjectives and nouns are modeled.

The Topic Extractor then selects the topics: if both the bigram words are labeled as proper noun or common noun, then the first noun word of the bigram is selected as label. If a pair composed by adjective and noun is selected, then it’is reported as topic. There are many different cases and, to the basic algorithm, also verbs have been applied to the Context Free Grammar.

In particular the interactions between verbs and nouns are currently under test, because it should be possible to isolate future tense sentences by modeling verbs and modals in the Context Free Grammar.

*Test*

Two articles have been randomly picked up for Proof-Of-Concept purposes: articles a00124022 and a00124023 have been choosen. For test purposes the algorithm is constrained to just the two chosen articles, but by setting *‘tests = articles’* it works on the whole articles corpus.

Starting from *‘amis\_articles\_27\_11\_2016\_indexed.jsonl’* the algorithm firstly breaks all the articles in different sentences (relying on *‘SBD.py’)*, then the function *‘sentence\_by\_sentence()’* in *‘sentence\_extraction\_POC.py’* extracts the topics (relying on *‘sentence\_tagger.py’*).

As last step a dictionary is created, where the keys are the sentences and the values are the topics, and it’s writen in the JSON file called *‘amis\_articles\_27\_11\_2016\_sentences\_from\_indexed.jsonl’.*

Test results are:

* a00124022

u'Plans for privatizing Serbias bloated state-owned companies are likely to feature extensive restructuring and large scale redundancies.': [u'Plans', u'Serbias', u'large scale redundancies'], u'With a fresh mandate, painful policies will be easier for Vucic to implement,Miha Hribernik, Senior Europe Analyst at Verisk Maplecroft, said by e-mail.': [u'fresh mandate', u'painful policies', u'Vucic', u'Miha Hribernik', u'Senior Europe Analyst', u'Verisk Maplecroft'], u'Vucic hasnt named a potential coalition partner, saying before the ballot only that hed reject any party that wouldnt agree to a pro-EU path.': [u'Vucic', u'potential coalition partner', u'pro-EU path'], u'Serbian Prime Minister Aleksandar Vucic said his Progressive party would form a new pro-European government at the end of next month, capitalizing on a snap election victory that has handed him four more years in power.': [u'Serbian Prime', u'Aleksandar Vucic', u'Progressive', u'new pro-European government', u'snap election victory'], u'The Progressives, who forced a ballot two years early for the second time since 2012, took 48.2 percent of Sundays vote, the Election Commission said in the capital, Belgrade, citing results with 97 percent of ballots counted.': [u'Progressives', u'Sundays', u'Election Commission said', u'Belgrade'], u'Fitch revised Serbias B+ credit rating to positive from stable in December.': [u'Fitch', u'Serbias B+', u'December'], u'Dacic has embraced the prime ministers European focus reluctantly and has favored stronger ties with Russia, which may encourage Vucic to turn to one of the smaller pro-EU groups or even poach lawmakers from rival voting coalitions.': [u'Dacic', u'prime ministers', u'European focus', u'Russia', u'Vucic', u'pro-EU groups', u'poach lawmakers'], u'The Socialist Party, led by Foreign Minister Ivica Dacic, was second with 11 percent, or 29 seats in the assembly.The Progressives also overcame resurgent nationalists, including acquitted war-crimes suspect Vojislav Seselj, whose Radical Party was third with 8.1 percent, or 22 seats.': [u'Socialist Party', u'Foreign Minister', u'Ivica Dacic', u'Progressives', u'resurgent nationalists', u'Vojislav Seselj', u'Radical Party'], u'Potential candidates may include the Socialists, with whom he has shared power since 2012.': [u'Potential candidates', u'Socialists'], u'The bespectacled 46-year-old, once a close ally of war-time leader Slobodan Milosevic, has vowed to end state support of unprofitable companies, shrink the public administration, and liberalize markets in the country of 7.2 million people.': [u'bespectacled 46-year-old', u'close ally', u'war-time leader', u'Slobodan Milosevic', u'state support', u'unprofitable companies', u'public administration', u'liberalize markets'], u'Vucic would prefer sharing the reform burden with other coalition partners, Dan Bucsa, lead central and eastern Europe economist at UniCredit Bank AG in London, said by e-mail.': [u'Vucic', u'reform burden', u'coalition partners', u'Dan Bucsa', u'Europe', u'UniCredit', u'AG', u'London'], u'The Progressives will hold a congress on May 28, and we expect that a new cabinet will be formed immediately after that, the premier told reporters in Belgrade on Monday.': [u'Progressives', u'May', u'new cabinet', u'Belgrade'], u'Four other parties also made it into parliament after 56.3 percent of voters cast ballots.': [u'voters cast ballots'], u'The dinar gained almost 0.1 percent to trade at 122.69 per euro.': [], u'Reforms may take a toll on the governments popularity, but a four-year mandate is long enough to ensure that higher growth and a rebound in revenue would occur before the next parliamentary elections are held.': [u'Reforms', u'governments popularity', u'four-year mandate', u'parliamentary elections'], u'Vucics party won 131 of parliaments 250 seats, a smaller majority than it achieved two years ago.': [u'Vucics'], u'Future measures taken to cut public debt and the current-account gap could be seen as ratings positive, Richard Grieveson, an analyst at Fitch Ratings, said in a note.': [u'public debt', u'current-account gap', u'Richard Grieveson', u'Fitch Ratings'], u'Before the vote, he pledged to form a coalition government to continue a painful economic overhaul endorsed by the International Monetary Fund.': [u'coalition government', u'painful economic overhaul', u'Monetary'], u'The yield on Serbian dollar bonds maturing in 2021 fell three basis points to 4.537 percent at 1:03 p.m. in Belgrade, according to data compiled by Bloomberg.': [u'Serbian', u'dollar bonds maturing', u'basis points', u'Belgrade', u'Bloomberg'], u'Bucking a trend of growing anti-EU sentiment in other eastern European states including Poland and Hungary, Vucic pledged to keep Serbia on a path to be prepared to join the bloc in 2020.': [u'Bucking', u'anti-EU sentiment', u'European states', u'Poland', u'Hungary', u'Vucic', u'Serbia'], u'Almost two decades after the bloody wars that tore apart former Yugoslavia, Serbia is one of Europes last ex-communist nations to embark on a wide-scale overhaul of its economy, long after others including the Czech Republic and neighboring Hungary.': [u'bloody wars', u'Yugoslavia', u'Serbia', u'Europes', u'ex-communist nations', u'wide-scale overhaul', u'Czech', u'Hungary'], u'Its living standards also lag those of richer EU states, including former Yugoslav partners Slovenia and Croatia, which joined the bloc in 2004 and 2013.': [u'EU', u'Yugoslav', u'Slovenia', u'Croatia'], u'While the measures may trigger job losses, the IMF says theyre necessary to heal an economy suffering from unemployment that exceeds 18 percent and an average take-home wage of $407 a month.': [u'trigger job losses', u'IMF', u'average take-home wage'], u'Vucic said his new government would stick with its $1.2 billion precautionary agreement with the IMF, and it would continue to rid the public balance sheet of more than 500 money-losing state-owned companies that sap as much as $1 billion from the budget a year.': [u'Vucic', u'new government', u'precautionary agreement', u'IMF', u'public balance sheet']

* a00124023

u'Higher disposable incomes arent translating into increased retail spending.': [u'disposable incomes arent translating'], u'The investor group that acquired it tried unsuccessfully to raise the money needed to meet contractual payments.': [u'investor group', u'contractual payments'], u'Next Chief Executive Officer Simon Wolfson likened the year ahead to walking up the down escalator.': [u'Executive Officer', u'Simon Wolfson'], u'When the weather improves, we hope consumers will come back.': [u'weather improves'], u'While BHS had a dowdy image, even some high-street retailers with a more contemporary profile have struggled lately.': [u'BHS', u'dowdy image', u'high-street retailers', u'contemporary profile'], u'The number of shoppers on British high streets declined 3.9 percent in March, according to data from researcher Springboard.': [u'British high streets declined', u'March', u'Springboard'], u'The move followed the sale of the company for a mere one pound last year by billionaire retail entrepreneur Philip Green.': [u'retail entrepreneur', u'Philip Green'], u'The U.K. has been tough, AB Foods Chief Executive Officer George Weston said in an interview, citing the recent impact on sales of unseasonably cool temperatures.': [u'U.K.', u'AB', u'Executive Officer', u'George Weston', u'recent impact', u'cool temperatures'], u'Britains high-street retailers are feeling low.': [u'Britains', u'high-street retailers'], u'A competitor, Next Plc, lowered its forecast for growth in sales of goods at full price by 2 percentage points in the current financial year, a reversal from years of outperformance.': [u'Plc', u'full price', u'percentage points', u'current financial year'], u'Though BHS stores remain open for now, if the chain were to shut down it would be the biggest retail failure in Britain since the demise of Woolworths Group Plc in 2008.': [u'BHS', u'retail failure', u'Britain', u'Woolworths', u'Plc'], u'The companies -- department store operator and formalwear retailer Austin Reed -- are the latest victims of a deepening crisis, as shoppers gravitate instead to out-of-town malls or online retailers.': [u'department store operator', u'formalwear retailer', u'Austin Reed', u'shoppers gravitate', u'out-of-town malls', u'online retailers'], u'Austin Reed, a 116-year-old suit seller, appointed AlixPartners as its administrator Tuesday, citing cash flow difficulties arising from challenging retail market conditions.': [u'Reed', u'116-year-old suit seller', u'AlixPartners', u'cash flow difficulties arising', u'challenging retail market conditions'], u'Traditional retail businesses with high-street stores have high fixed costs, Alan Stewart, chief financial officer of Tesco Plc, said in a wide-ranging interview.': [u'Traditional retail businesses', u'high-street stores', u'Alan Stewart', u'financial officer', u'Tesco Plc', u'wide-ranging interview'], u'The decline in traffic is a continuation of a longer-term trend caused by on-going structural change within the retail industry, Helen Dickinson, chief executive of industry lobby group the British Retail Consortium, said by email.': [u'longer-term trend caused', u'retail industry', u'Helen Dickinson', u'chief executive', u'industry lobby group', u'Consortium'], u'U.K. same-store sales at Primark, a discount fashion chain owned by Associated British Foods Plc, fell about 1 percent in the most recent quarter.': [u'U.K.', u'same-store sales', u'Primark', u'discount fashion chain owned', u'Plc', u'recent quarter'], u'The high street is a complex urban ecosystem that has shown tremendous capacity to adapt, and it will continue to do so, he said.': [u'high street', u'complex urban ecosystem', u'tremendous capacity'], u'Since then, the country has lost several other longtime high-street chains, including HMV music stores, Comet electronics shops and Jessops photo centers.': [u'longtime high-street chains', u'HMV', u'music stores', u'Comet', u'electronics shops', u'Jessops', u'photo centers'], u'Green was unable to turn around BHS, which he owned for about 15 years, despite his success with other British retailers such as apparel chain Topshop.': [u'BHS', u'British retailers', u'apparel chain', u'Topshop'], u'If their business starts changing and they dont have a capital structure thats sufficiently strong to withstand that change then theyre in a weakened position.': [u'capital structure thats'], u'The move jeopardizes 1,184 jobs.': [u'move jeopardizes'], u'In the last week alone, two stalwarts of the countrys downtown shopping districts have moved to appoint administrators, similar to filing for Chapter 11 bankruptcy protection in the U.S.': [u'bankruptcy protection', u'U.S'], u'Analysts say BHS, a nearly nine-decade old chain of 164 stores selling everything from lighting to clothing, failed to adapt its product line as well as its business model.': [u'BHS', u'old chain', u'product line', u'business model'], u'BHS appointed administrators Monday, putting 11,000 jobs at risk.': [u'BHS'], u'In the U.K., 13.4 percent of all retail sales took place online in 2015, compared with 9.2 percent in the U.S. and a global average of 7.3 percent, according to researcher Euromonitor.': [u'U.K.', u'retail sales took place online', u'U.S.', u'global average', u'Euromonitor'], u'He said he saw signs of life in the artisanal bakers and butchers that are returning to some of these shopping districts in affluent regions in and around London.': [u'artisanal bakers', u'affluent regions', u'London'], u'Bricks-and-mortar shops have been undermined by Britains leading position in e-commerce.': [u'Bricks-and-mortar', u'Britains'], u'Long-time occupants of prime retail space are moving out, and in many cases theyre being replaced by discount chains and thrift stores.': [u'Long-time', u'prime retail space', u'cases theyre', u'discount chains', u'thrift stores'], u'The decades-old concept is now being upended by changing consumer behavior and the inability of some retailers to remain relevant, a fate that contributed to BHSs downfall.': [u'decades-old concept', u'consumer behavior', u'BHSs'], u'Customers dont differentiate between buying online, on a mobile device or in-store.': [u'Customers', u'dont differentiate', u'mobile device'], u'It now sees revenue falling as much as 1 percent or gaining 4 percent at best, compared with a Jan. 5 forecast for growth of 1 percent to 6 percent.': [u'Jan.'], u'Unless retailers adapt their business to meet the needs of todays consumer, we will see more high-profile casualties, as the combination of fragile economic conditions and the disruptive force of online retailing continues to bite, Jason Shorrock, vice president of EMEA retail strategy at consultancy JDA, said by email.': [u'Unless', u'retailers adapt', u'todays consumer', u'high-profile casualties', u'economic conditions', u'disruptive force', u'Jason Shorrock', u'vice president', u'EMEA', u'retail strategy', u'JDA'], u'Wolfson cited the likelihood that slowing growth in disposable incomes would hamper spending.': [u'Wolfson', u'disposable incomes', u'hamper spending'], u'The slump is changing the fabric of a British institution that draws masses of shoppers to town centers on weekends with a collection of brands ranging from John Lewis department stores to booksellers like Waterstones to apparel chains such as Marks & Spencer and Next.': [u'British institution', u'draws masses', u'town centers', u'John Lewis', u'department stores', u'Waterstones', u'apparel chains', u'Marks', u'Spencer'], u'The decline was the first in more than a decade, according to analysts.': [], u'Neil Wrigley, professor of human geography at the University of Southampton, said the high street will survive despite changing consumer behavior and missteps by individual retailers.': [u'Neil Wrigley', u'human geography', u'Southampton', u'high street', u'consumer behavior', u'individual retailers'], u'fell 1.3 percent in March, according to the Office for National Statistics.': [u'March', u'Statistics']

The results show that in the vast majority of the cases the main topics, namely sentence subject and/or context, are reflected at least in one of the same sentence extracted topic-keywords. There are obviously a few errors, but the impact has still to be evaluated in terms of sentiment extraction, namely the comparison between old sentiment-price correlation.

In order to have further evidences about methodology goodness, a wheat related article has been randomly chosen (from old *‘wheat\_articles.txt’* file, id: 96159). The results are the following:

u'The fall primarily reflected a six-and-a-half year low in international palm oil prices, mainly the result of slowing import demand, notably by India and China, amid expectations of rising production.': [u'six-and-a-half year', u'international palm oil prices', u'import demand', u'India', u'China'], u'Read the full brief, ': [u'Read'], u'Meanwhile, the latest FAO forecast for global cereal production in 2015 stands at 2.54 billion tonnes, 13.8 million tonnes more than expected in July, but still 21 million tonnes (0.8%) below the 2014 record.': [u'FAO', u'global cereal production', u'July'], u'Based on the current forecast of 501 million tonnes, global rice production would be 3.6 million tonnes (0.7%) greater than in 2014.': [u'Based', u'global rice production'], u'The revision was driven by higher expectations for crops in Australia, the E.U., Russia and Ukraine, more than offsetting a lower production forecast for Canada, where major growing areas continued to be affected by dry conditions.': [u'Australia', u'E.U.', u'Russia', u'Ukraine', u'Canada', u'dry conditions'], u'The trade-weighted FAO Food Price Index tracks international market prices for five major food commodity groups: cereals, meat, dairy products, vegetable oils and sugar.': [u'FAO', u'Price Index', u'international market prices', u'major food commodity groups', u'dairy products', u'vegetable oils'], u"The FAO forecast for world cereal stocks by the close of the seasons in 2016 has been raised by almost 12 million tonnes since July, to 643 million tonnes, on account of more optimistic crop prospects, putting this season's ending inventories only 6 million tonnes (1%) below its record high opening level.": [u'FAO', u'cereal stocks', u'July', u'optimistic crop prospects', u"season 's ending inventories", u'record high opening level'], u'Continued improvements in production prospects for 2015-16 were largely behind the cereal price slides.': [u'Continued', u'production prospects', u'cereal price slides'], u'The index averaged 155.7 points in August 2015, down 5.2% from July, the steepest monthly drop since December 2008, with virtually all major food commodities registering marked dips.': [u'August', u'July', u'monthly drop', u'December', u'major food commodities registering'], u'However, this remains 19.9 million tonnes (1.5%) short of the 2014 record.': [], u"ROME, ITALY International food commodity prices continued to decline in August as ample supplies, a slump in energy prices and concerns over China's economic slowdown all contributed to the sharpest fall of the Food and Agriculture of the United Nations (FAO) Food Price Index in almost seven years.": [u'ROME', u'ITALY', u'International food commodity prices continued', u'August', u'ample supplies', u'energy prices', u'China', u'economic slowdown', u'sharpest fall', u'FAO', u'Price Index'], u'The upgrade from July was mostly driven by better than anticipated growing conditions in Argentina (maize), Brazil (maize) and the U.S. (maize and sorghum), which more than offset a drop in maize production in the E.U., where dry and hot weather dampened yield expectations.': [u'July', u'Argentina', u'Brazil', u'U.S.', u'maize production', u'E.U.', u'hot weather', u'yield expectations'], u'A new FAO Cereal Supply and Demand Brief, also issued on Sept. 10, raises forecasts for global coarse grains output by 7.5 million tonnes to 1.311 billion tonnes.': [u'FAO', u'Cereal Supply', u'Demand Brief', u'Sept.', u'raises forecasts', u'global coarse grains output'], u'In August, the cereal price index averaged 154.9 points, down 7% from July and 15.1% from last year a decline driven by falling wheat and maize prices that reversed two consecutive months of modest increases.': [u'August', u'cereal price index averaged', u'July', u'maize prices', u'consecutive months', u'modest increases'], u'Prospects for world rice production (milled basis) have also improved since July 2015, albeit by only 1.3 million tonnes, mainly attributable to India, where plantings are progressing in line with last season in spite of the prevailing El Nio.': [u'Prospects', u'world rice production', u'July', u'India', u'El Nio'], u'The upward revision resulted from more buoyant production prospects for coarse grains, wheat and rice.': [u'buoyant production prospects', u'coarse grains'], u'The vegetable oil price index averaged 134.9 points in August, down 8.6% from July, and its lowest level since March 2009.': [u'vegetable oil price index averaged', u'August', u'July', u'March'], u'As the harvest is nearing completion in the Northern Hemisphere, the global wheat production forecast for 2015 is becoming firmer, with 728 million tonnes now expected, 5 million tonnes more than previously foreseen.': [u'Hemisphere', u'global wheat production']

Also in this case most of the sentences have keywords which reflect the sentences main topics.

An interesting error is the following:

*'The upward revision resulted from more buoyant production prospects for coarse grains, wheat and rice.': [u'buoyant production prospects', u'coarse grains']*

In this case wheat and rice are not picked as topic-keywords because the bigrams *[wheat, and]* and *[and, rice]* are not considered relevant, as, in general, all couples *[article, noun]* and *[noun, article]*. This situation is very similar to this paper first showed example. The error is given by the fact that this sentence relates to all coarse grains, rice and wheat (expectations), but only coarse grains appears as topic-keyword.

In a few words the algorithm is not flawless, but it can be taken as a good approximation of reality.

**Conclusion**

As showed in this Proof Of Concept, the Sentence Topic extractor can be used to compute a set of keywords per each article sentence. Such keywords are quite descriptive of each sentence topic, namely subject and context.

Given the fact that in sentiment extraction it’s important to analyze sentences related to just one entity, this sentence topic extraction method allows to pick up sentences by topic-keywords which are more related to the sentence major entity.

The main idea is to select sentences by looking to these topic-keywords rather than just the simple words, namely to check if the commodity keywords sets extracted by Michael Kao are present in each sentence topic-keywords set. In this way it should be possible on one hand to build a sentiment index per each commodity, while on the other hand to differentiate between different aspects of the same commodity using the others same sentence topic-keywords. This last objective could be reached by looking for pairs of keywords in topic-keywords, for example *“Rice, Import”.*

Moreover it should be possible to select just future sentences by changing the structure of the Context Free Grammar: for example, isolation of future tense sentences can be possible by identifying the *“will/could/would + verb”* structure by modeling in the CFG as relevant the pair *[Modal, Verb]* and the other future tense costruction *“is going to”*.