Project Information

Title

Turkish Summarization Using Deep Learning

Target

Public [✓]   Restricted [ ]

Proposer Information

Name(s)	Fatih Mehmet Güler – PragmaCraft Dr. Ayşenur Birtürk (Supervisor)
E-Mail(s)	fmguler@pragmacraft.com birturk@ceng.metu.edu.tr

IP (Intellectual Property) Information

Intellectual Property rights belong to PragmaCraft Yazılım Ltd. Şti.

Project Description and Background Information

Description

The project aims to summarize any Turkish web page content, using deep learning and word vectors. The end product is a browser plugin/extension. In supported browsers, user clicks the browser extension button named 'Summarize' or 'TL;DR'. Then the extension reads the page contents, sends it to the server, then the server side application converts text to corresponding word vectors, and creates clusters using k-means (or similar) clustering algorithms and returns the client (browser extension) resulting summary of the text.Browser extension displays the summary text to user or removes non summary text from the page leaving only the summary.Another use case is searching the page with a keyword, e.g. keyword based summarization.

Similar Products/Projects

(Search and identify similar products and/or projects, and provide brief information about them.)
There are bunch of summarizers for English language (smmry.com, autosummarizer.com, freesummarizer.com etc) and browser extensions (TLDR, SummarizeThis, Summarizer, etc) but none for Turkish. Furthermore we believe that existing summarizers don't use deep learning.

-smmry.com

It summarize the text given as URL or user file input. Number of the sentences in summary can be decided by the user. It gives the reduction percentage of the summary as information. This reduction percentage can be up to %95. In this website, user can summarize the text according to keywords.

-TLDR Google Chrome Extension

This extension works on Google Chrome web browser. When website of an article is open on browser, if the user clicks on the extension button, TLDR serves four different options of summarized article. The user can see either %10, %15, %50 and %85 shortened version of the article as a summary.

Justification of the proposal

(Describe the purpose of the project.)
The purpose of the project is to make use of existing unsupervised learning techniques for natural language processing in Turkish.
(Why is there a need to develop the project you are proposing?)
There is very little research on Turkish NLP with deep learning, and the use cases serve a real need - information overload.
(Which basic problems does your project aims to solve?)
Facilitating the reading of Turkish articles with summarizing them meaningfully.

Contributions, Innovation and Originality Aspects of the Project

(State innovation and originality aspects as well as contributions planned in the project.)

To our knowledge, deep learning has never used in a Turkish NLP project. We aim to contribute to Turkish NLP society in terms of deep learning know how, creating word vectors for Turkish, summarization using word vectors.

(If there exists any developed products or existing projects similar to yours that target the same problem area of your proposal, how will your targeted product be different, possibly be better than those existing ones?)

Existing summarizers do not employ deep learning methods and most of them are not integrated to browsers as browser extensions to improve user experience.

(What will be the advantages and distinctive characteristics of your targeted product?)

The most significant advantage is better summarization performance, since unknown words and word similarities are better handled by word vectors. Moreover seamless integration with browser contributes to a better user experience.

(What are the contributions of your project to technological development at national and international levels?)

Turkish word vectors will contribute to better coverage in NLP applications , and summarization with word vectors contribute to NLP deep learning community.

(Does your project have any potential to initiate further research and/or development activities in the same or different technological areas?)

Word vectors are the basis for deep learning with NLP. Having pre-trained word vectors can lead to bunch of other NLP research using deep learning, such as sentiment analysis, machine translation, question answering, parsing, relationship extraction, etc.

Technical Aspects of the Project

(Provide some technical elaboration of the project -not as detailed as an SRS or an SDD, but detailed enough to visualize the technical aspects of the finished project as closely as possible.)

Client side (browser extension)

Lives in the users' browser and collects the text in the web page to be summarized.

Serve side (web application)

Receives the text to be summarized, summarizes it with pre-trained word vectors and returns the summary to client.

Word vectors

A separate phase of the project is to train word vectors using a large corpus. Word2vec or Glove tools to be used. These tools create a language model and assign an n dimensional vector to each unique word. The cosine distance between related words are smaller, and unrelated words are bigger.

Targeted Output, Targeted User/Domain Profile

(Describe what the end-product(s) will look like and how it will be used.)

End product is a browser extension. User clicks this summarize button in the browser and receives a summary.

(Provide tangible success measures and goals.)

Extracting a meaningful summarization of Turkish articles that is also easily comprehensible for Turkish speakers. Providing a high level user experience.

(Provide information about the users / user groups and/or domain that will utilize the product.)

Knowledge workers, who has to read a lot of information on the web, students, researchers, software developers, lawyers, senior executives, etc.

Project Development Environment

(State the planned hardware / software technologies and programming languages to be used.)

For browser extension and server application Javascript, Java respectively. For language modeling word2vec tool (C), Glove (C), for deep learning Torch (Lua). For deep learning we plan to utilize GPU.

(State the planned methods, tools and techniques to be used.)

Training word vectors with Word2vec, Web application with Java & Spring, Chrome extension with Javascript.

External Support

(List any required hardware and software support for your project.)

For word vector training, we may need a GPU with NVIDIA CUDA support, for faster training. Word2vec, Glove for word vectors. Torch for deep learning.

(List and describe the resources provided by a sponsor -if any.)

PragmaCraft will provide a deep learning environment with GPU (remotely).

(Do you plan to utilize external support including know-how, consultancy services, etc. for some minor parts of the project?)

PragmaCraft will provide the know-how for deep learning, web application development, browser extension development.

References

(Please provide references / links (URLs) for your answers in above sections.)

Word vectors;
https://code.google.com/p/word2vec/
http://nlp.stanford.edu/projects/glove/

Deep Learning;
http://torch.ch/
https://en.wikipedia.org/wiki/Torch_(machine_learning)

https://developer.nvidia.com/about-cuda

Existing summarization tools;
autosummarizer.com
autosummarizer.com/
freesummarizer.com/
https://chrome.google.com/webstore/search/summarize