2. **【現在までの研究状況】**(図表を含めてもよいので、わかりやすく記述してください。様式の変更・追加は不可(以下同様))

① これまでの研究の背景、問題点、解決方策、研究目的、研究方法、特色と独創的な点について当該分野の重要文献を挙げて記述してください。

② 申請者のこれまでの研究経過及び得られた結果について、問題点を含め①で記載したことと関連づけて説明してください。なお、これまでの研究結果を論文あるいは学会等で発表している場合には、申請者が担当した部分を明らかにして、それらの内容を記述してください。

❖ Research Background

Large-scale knowledge bases like Freebase and DBpedia, consist of a large pool of information with real-world entities as nodes and their relations as edges. Each directed edge, along with its head entity and tail entity, constitute a triple, i.e., (head entity, relation, tail entity), which is also named as a fact. Because of their large volume and complex data structures, it is difficult for non-technical users to access the substantial and valuable knowledge in them. To bridge this gap, Question answering over knowledge base (KB) aims at providing a way to automatically translate the end users natural language questions into structured queries and returning entities and/or predicates from the KB as answers hence allowing non-technical users to access the information they want. For example, given a question ”where was Barack Obama born?”, Question answering over KB aims at identifying its corresponding triple, i.e.,(Barack Obama, people/person/place/of/birth, Honolulu).

❖ Research Purpose, Objectives, methods and Contribution

In this work, we propose an approach for efficient question answering (QA) of simple queries over knowledge base (KB). Our approach is different from previous end-to-end complex models, we take a simple modular approach of decomposing the simple question-answering task in three different components, as shown in figure 1:

(1). Entity detection, where standard Recurrent Neural Network (RNN), and Conditional Random Field (CRF) are applied to identify entities in the question,

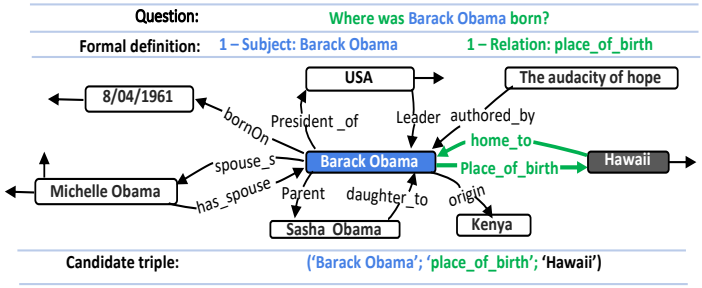

Figure 1: The question answering process

(2). We then link the identified entities to there corresponding nodes in the KB using an inverted index to come up with a candidate list with their respective score.

(3). Relation prediction where a question is classified as one of the relation types in the KB, we apply standard (RNN) plus standard Convolutionl Neural Network (CNN) to do this.

With the objective of enhancing the performance of the simple question answering system using baseline methods, we examine the necessity of complex models for the simple question answering task as applied by previous related work, and we do this by exploring the performance of baseline methods both standard neural network and non neural network techniques that perform reasonably well on a similar task. Our work makes the following contributions;

(i) We propose a simple yet effective approach, our approach is faster/efficient to train the network and performs reasonably well compared to previous complex approaches on a similar task. (ii) We introduce a novel index that relies on the relation type to filter out subject entities from the candidate list so that the object entity with the highest score becomes the answer to the question .

❖ Research Progress and Obtained Results

(I). Entity detection:

To identify the entity in the question we formulate this as a sequence-labeling problem where each word or token is tagged as entity or non-entity, **I**: entity and **0**:non- entity. We apply both recurrent neural network (RNN) and conditional random field to this task. And generate candidate entities. Figure 2 on the right side shows the RNN architecture for entity detection.

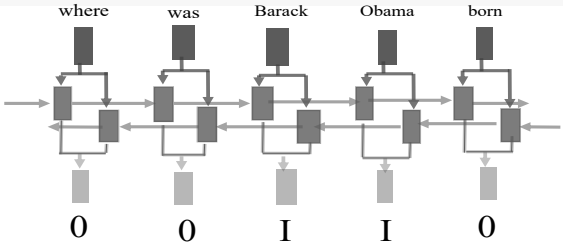

Figure 2: RNN architecture for entity detection

We conducted experiment using 100,000 questions from the simple questions dataset and freebase as the knowledge base. For entity detection, we evaluate the precision, recall and F1-score on the token span level. This means that the predicted entity token span exactly matches the ground truth (a true positive span). The results reveal that RNN perform better with F1-score of 92.5%. We also noticed that CRF result was 90.2% on the same task. Our results were presented to a domestic conference and we were awarded the best student presenter award.

申請者登録名　Buzaaba Happy

DC

¥（現在までの研究状況の続き）

(II). Entity linking:
We link the generated candidate entities to actual nodes in the freebase knowledge base. We use the inverted index to map any entity n-gram to all nodes in the knowledge base. And we compute this association of entity n-grams to nodes in the knowledge base by term frequency inverse document frequency (tf-idf) to come up with candidate entities and their respective score. For example, assuming a node referring to former United States president ”Barack Obama” exists in the knowledge base, the alias of this entity node will be a name with two uni-grams (”Barack”, ”Obama”) and a single bi-gram of the entire name (”Barack Obama”). We therefore index each of these n-grams with tf-idf weights. And the weights would be computed as follows;

I(”Barack”) → {node : ei, score: tf-idf(”Barack”, ”Barack Obama”)}
I(“Barack Obama”) → {node : ei, score: tf-idf(“Barack”, “Barack Obama”)}

We perform this for every n-gram of every entity node in the KB and we are able to generate a list of candidate entities with their associated scores.

(III). Relation Prediction:
We classify the question as one of the 1837 unique relation types in freebase knowledge base. Assuming we have a question ”how old is barack obama”, the relation type, which refers to the date of birth would be; ”people_person_bornOn”. We examine both recurrent and convolutional neural network for this task. Our precision results show that CNN outperforms RNN on this task with 81.92 and 81.55 respectively.

(IV). End-to-end Results: Our best model combination of entity detection and relation prediction achieves 74.64% accuracy. Our results outperform the complex neural network models like Bodes et al’s memory network, Golub and He’s attention-enhanced encoder-decoder framework and Lukovinikov et al’s complex character and word-level encoding.
Our results were presented to a domestic conference and we were awarded the best student presenter award.

[1] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale simple question answering with memory networks. arXiv:1506.02075, 2015.
[2] D. Golub and X. He. Character-level question answering with attention. arXiv:1604.00727, 2016.
[3] D. Lukovnikov, A. Fischer, J. Lehmann, and S. Auer. Neural network-based question answering over knowledge graphs on word and character level. WWW. 2017.
[4] O. Irsoy and C. Cardie. Opinion mining with deep recurrent neural networks. EMNLP 2014

3. 【これからの研究計画】

(1) 研究の背景

2.で述べた研究状況を踏まえ、これからの研究計画の背景、問題点、解決すべき点、着想に至った経緯等について参考文献を挙げて記入してください。

❖ Future work

In our previous work, we focus on answering simple questions, which require extracting a single fact of the form (subject, predicate, object) from the knowledge base to be answered. In the work we propose a simple modular approach for decompose the simple question-answering task in a three step-pipeline: entity detection, entity linking and relation prediction. Softer we have made the following contributions; (i) we propose a simple yet effective approach, our approach performs reasonably well compared to previous complex approaches on a similar task. (ii) We introduce a novel index that relies on the relation type to filter out subject entities from the candidate list so that the object entity with the highest score becomes the answer to the question. In the future work, we will examine the question-answering problem on much complex questions that might have more than one entity and multi-relation. Our current work uses freebase as the knowledge base, we believe that considering multiple knowledge sources in our future work would improve the accuracy of our results on both simple and complex questions.

❖ Importance of Future work

Moving forward, we are more interested in formally decomposing the simple question-answering task into sub-problems and solve each separately by exploring different methods for each task. We also strongly believe that conducting an in depth error analysis is equally important to understand why a certain level of accuracy is achieved.

Our previous/current work, focus only on simple questions with one subject and one relation. This may not be applied for questions with more than one subject and multi-relation types. In the future work we will; (i). Extend our work to more complex questions. One other problem that we faced was the incompleteness of the knowledge base. Some questions could not be answered because there were no triples in the knowledge base that provides an answer to the question. We will; (ii). Consider combination of multiple knowledge sources, which would improve the performance of our results.

[5] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, W. Cohen, Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text EMNLP 2018

(2) **研究目的・内容**（図表を含めてもよいので、わかりやすく記述してください。）

① 研究目的、研究方法、研究内容について記述してください。
② どのような計画で、何を、どこまで明らかにしようとするのか、具体的に記入してください。
③ 所属研究室の研究との関連において、申請者が担当する部分を明らかにしてください。
④ 研究計画の期間中に異なった研究機関（外国の研究機関等を含む。）において研究に従事することを予定している場合はその旨を記載してください。

❖ Research Purpose

The purpose of this study is to propose a more practical setting, namely QA over the combination of a KB and entity linked text, which is appropriate when an incomplete KB is available with a large text corpus. We propose a novel model, for extracting answers from a question-specific sub-graph containing text and KB entities and relations. A novel approach designed to operate over heterogeneous graphs of KB facts and text sentences is proposed Figure 3.
The proposed work builds upon recent work on graph representation learning (Kipf and Welling, 2016; Schlichtkrull et al., 2017), but propose two key modifications to adopt them for the task of QA.

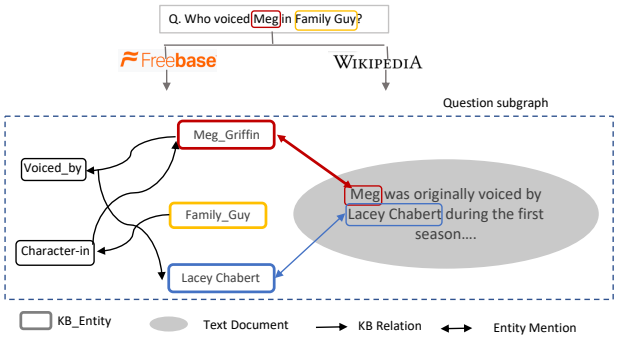

Figure 3: Heterogeneous graphs of KB and text

(i). First, we propose heterogeneous update rules that handle KB nodes differently from the text nodes: for instance, RNN-based updates will be used to propagate information into and out of text nodes.
(ii). Second, we will introduce a directed propagation method, inspired by personalized Pagerank in information retrieval (Haveliwala, 2002), which constrains the propagation of embeddings in the graph to follow paths starting from seed nodes linked to the question.

❖ Research Method:

1. **Description:**
The task is that given a natural language question; extract its answers from the knowledge graph. There may be multiple correct answers for a question. Our proposed approach assumes that the answers are entities from either the documents or the KB. We are interested in a wide range of settings, where the KB varies from highly incomplete to complete for answering the questions.
To solve this task, we proceed in two steps;
(I). We will extract a sub-graph, which contains the answer to the question with high probability. The goal for this step is to ensure high recall for answers.
(II). We will use our proposed method to learn node representations in the sub-graph conditioned on the question, and the node representations will be used to classify each node as being an answer or not.

1.1 Question Sub-graph Retrieval

We will retrieve the sub-graph using two parallel pipelines; One over the KB, which returns a set of entities, and the other over the text corpus, which returns a set of documents. The retrieved entities and documents will then be combined with entity links to produce a fully connected graph.
KB Retrieval. To retrieve relevant entities from the KB entity linking will be performed on the question to produce a set of seed entities. Next we will run the Personalized PageRank (PPR) method (Haveliwala, 2002) around these seeds to identify other entities, which might be an answer to the question. The edge-weights around the seed entities are distributed equally among all edges of the same type, and they are weighted such that edges relevant to the question receive a higher weight than those, which are not. Specifically, we will average word vectors to compute a relation vector from the surface form of the relation, and a question vector from the words in the question, and we will use cosine similarity between these as the edge weights. After running PPR we will retain the top entities by PPR score, along with any edges between them, and add them to the sub-graph.
Text Retrieval. We intend to use Wikipedia as the corpus and retrieve text at the sentence level, i.e. documents in the corpus are defined along sentences boundaries. Text retrieval will be performed in two steps: First we will consider the top 5 most relevant Wikipedia articles, using the weighted bag-of-words model from DrQA (Chen et al., 2017), we will populate a Lucene index with sentences from these articles, and retrieve the top ranking based on the words in the question. For the sentence-retrieval step, we intend to include the title of the article as an additional field in the Lucene index. As most sentences in an article talk about the title entity, this helps in retrieving relevant sentences that do not explicitly mention the entity in the question. We add the retrieved documents, along with any entities linked to them, to the sub-graph.
With this study, we will be able to achieve better results to enhance the performance of our question answering system that will allow non-technical users to pose natural language question on structured knowledge sources and access the information they need.

申請者登録名　Buzaaba Happy

DC

(3) 研究の特色・独創的な点

次の項目について記載してください。
① これまでの先行研究等があれば、それらと比較して、本研究の特色、着眼点、独創的な点
② 国内外の関連する研究の中での当該研究の位置づけ、意義
③ 本研究が完成したとき予想されるインパクト及び将来の見通し

❖ Distinctive Features of this study

Previous work attempts an early fusion strategy for QA over KB facts and text. Their approach is based on Key-Value Memory Networks (KV-MemNNs) coupled with a universal to populate a memory module with representations of KB triples and text snippets independently. The key limitation for this model is that it ignores the rich relational structure between the facts and text snippets. Our proposed method, on the other hand, will explicitly use this structure for the propagation of embeddings. We intend to compare the two approaches in our experiments, to evaluate the performance over all tasks.

❖ Significance of this study

The proposed approach is motivated from the large body of work on graph representation learning (Scarselli et al., 2009; Li et al., 2016; Kipf and Welling, 2016; Atwood and Towsley, 2016; Schlichtkrull et al., 2017). Like most other graph-based models, our model can also be viewed as an instantiation of the Message Passing Neural Network (MPNN) framework of Gilmer et al. (2017). Our approach can also be seen to be an inductive representation learners like GraphSAGE (Hamilton et al., 2017), but operate on a heterogeneous mixture of nodes and use retrieval for getting a sub-graph instead of random sampling.

[6] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner, G. Monfardini. The graph neural network model. IEEE 2009
[7] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel. Gated graph sequence neural networks. ICLR 2016
[8] T. Kipf, M. Welling. Semisupervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016
[9] J. Atwood, D. Towsley. Diffusion convolutional neural networks. NIPS 2016

(4) 年次計画

申請時点から採用までの準備状況を踏まえ、DC1 申請者は１～３年目、DC2 申請者は１～２年目について、年次毎に記載してください。元の枠に収まっていれば、年次毎の配分は変更して構いません。

（申請時点から採用までの準備）

Data in knowledge bases like Freebase and DBpedia, consist of a large pool of information with real-world entities as nodes and their relations as edges. Each directed edge, along with its head entity and tail entity, constitute a triple, i.e., (head entity, relation, tail entity), which is also named as a fact. It is difficult however, for non-technical users to access the substantial and valuable knowledge in these knowledge graphs because of their complex structure. Our work aims at providing a way to allow end users to pose natural language questions on knowledge graphs and access information they want. We are more interested in formally decomposing the question-answering task into sub-problems of entity detection, entity linking and relation prediction and solve each separately by exploring different methods for each task and compare with previous work for each particular problem.

（１年目）

A Scheme for Factoid Question Answering over Knowledge Base.
In this work we focus on answering simple questions. These are questions that have one subject and one relation. These questions can be answered by extracting a single triple from the KB to be answered. We decompose the simple QA task in a three step-pipeline: entity detection, entity linking and relation prediction. More precisely, our proposed approach is quite simple but performs reasonably well compared to previous complex approaches. We introduce a novel index that relies on the relation type to filter out subject entities from the candidate list so that the object entity with the highest score becomes the answer to the question. Furthermore, due to its simplicity, our approach can significantly reduce the training time compared to other comparative approaches. The experiment on the SimpleQuestions data set finds that basic LSTMs, GRUs, and non-neural network techniques achieve reasonable performance while providing an opportunity to understand the problem structure.

（２年目）

A modular Approach for Efficient Question Answering Over Knowledge Base.
Our current work focuses on answering complex natural language questions; these are questions with more than one entity and multiple-relations. We also are setting up to use multiple knowledge sources so as to overcome the challenge of KB base incompleteness faced in the previous work. The purpose of this study is to propose a more practical setting, namely complex QA over the combination of a KB and entity linked text, which is appropriate when an incomplete KB is available with a large text corpus. We propose a novel model, for extracting answers from a question-specific sub-graph containing text and KB entities and relations.

申請者登録名　Buzaaba Happy

Research Plan

Courses	2018		2019		2020	
	Spring	Autumn	Spring	Autumn	Spring	Autumn
1. Topic Description						
Complete required credits						
Conference artical1						
Journal artical1						
2. Making more surveys						
Data collection						
Apply new algorit ms						
Conference artical2						
Journal artical2						
3. Preparing Thesis and defense						

(5) 人権の保護及び法令等の遵守への対応

本欄には、研究計画を遂行するにあたって、相手方の同意・協力を必要とする研究、個人情報の取り扱いの配慮を必要とする研究、生命倫理・安全対策に対する取組を必要とする研究など法令等に基づく手続きが必要な研究が含まれている場合に、どのような対策と措置を講じるのか記述してください。**例えば、個人情報を伴うアンケート調査・インタビュー調査、国内外の文化遺産の調査等、提供を受けた試料の使用、侵襲性を伴う研究、ヒト遺伝子解析研究、遺伝子組換え実験、動物実験など**、研究機関内外の情報委員会や倫理委員会等における承認手続きが必要となる調査・研究・実験などが対象となりますので手続きの状況も具体的に記述してください。

なお、該当しない場合には、その旨記述してください。

N/A

4.【研究成果】(下記の項目について申請者が中心的な役割を果たしたもののみ項目に区分して記載してください。その際、通し番号を付すこととし、該当がない項目は「なし」と記載してください。申請者にアンダーラインを付してください。論文数・学会発表等の回数が多くて記載しきれない場合には、主要なものを抜粋し、各項目の最後に「他○報」等と記載してください。査読中・投稿中のものは除く)

(1) 学術雑誌等（紀要・論文集等も含む）に発表した論文、著書（査読の有無を区分して記載してください。査読のある場合、印刷済及び採録決定済のものに限ります。）
著者（申請者を含む全員の氏名（最大 20 名程度）を、論文と同一の順番で記載してください。）、題名、掲載誌名、発行所、巻号、pp 開始頁－最終頁、発行年をこの順で記入してください。

(2) 学術雑誌等又は商業誌における解説、総説

(3) 国際会議における発表（口頭・ポスターの別、査読の有無を区分して記載してください。）
著者（申請者を含む全員の氏名（最大 20 名程度）を、論文等と同一の順番で記載してください。）、題名、発表した学会名、論文等の番号、場所、月・年を記載してください。発表者に○印を付してください。（発表予定のものは除く。ただし、発表申し込みが受理されたものは記載しても構いません。）

(4) 国内学会・シンポジウム等における発表
(3)と同様に記載してください。

(5) 特許等（申請中、公開中、取得を明記してください。ただし、申請中のもので詳細を記述できない場合は概要のみの記述で構いません。）

(6) その他（受賞歴等）

(1) 学術雑誌等（紀要・論文集等も含む）に発表した論文、著書, (Articles published in academic Journal)
(No peer review)

1. Happy Buzaaba, “Monitoring Social and Economic characteristics from call detail records”, IPSJ SIG Technical Report, Information Processing Society of Japan, Vol.2017-IFAT-125, Article No. 7, pp.1-7, 2017.
2. Happy Buzaaba, Chieko Nakabasami, “Monitoring Social-Economic characteristics from call detail records”, IPSJ SIG Technical Report, Information Processing Society of Japan, Vol.2017-DC-105, Article No. 3,pp.1-7, 2017.

(2) 学術雑誌等又は商業誌における解説、総説
N/A

(3) 国際会議における発表
N/A

(4) 国内学会・シンポジウム等における発表
（口頭発表，査読なし)

3. **○Happy Buzaaba, “Monitoring Social and Economic characteristics from call detail records”, ” 第 125 回情報基礎とアクセス技術, 東洋大学, 2017 年 3 月.**
4. **○Happy Buzaaba, Chieko Nakabasami, “Monitoring Social-Economic characteristics from call detail records”, 第 105 回ドキュメントコミュニケーション研究発表, 凸版印刷(株)印 刷博物館,文京区, 東京都, 2017 年 7 月.**
5. **○Happy Buzaaba, Toshiyuki Amagasa, “A scheme for Factoid Question Answering over knowledge base” 第11回データ工学と情報マネジメントに関するフォーラム（第17回日本データベース学会年次大会）**、ホテルオークラJRハウステンボス 〒859-3296 長崎県佐世保市ハウステンボス町10番, **2019年3月4日-6日**

シンポジウム等における発表 **(ポスター)**

6. **○Happy Buzaaba, Toshiyuki Amagasa, “A scheme for Factoid Question Answering over knowledge base” 第11回データ工学と情報マネジメントに関するフォーラム（2019年3月4日-6日）**

(5) 特許等
N/A

(6) その他（受賞歴等）

7. Buzaaba Happy, Toyo University Graduate Schools of Regional Development studies Best masters Thesis award Fall 2017
8. Buzaaba Happy, 第 11回データ工学と情報マネジメントに関するフォーラム, 学生プレゼンテーション賞, 2019 年 3 月

申請者登録名　Buzaaba Happy

5.【研究者を志望する動機、目指す研究者像、自己の長所等】

日本学術振興会特別研究員制度は、我が国の学術研究の将来を担う創造性に富んだ研究者の養成・確保に資することを目的としています。この目的に鑑み、申請者本人の研究者としての資質、研究計画遂行能力を評価するために以下の事項をそれぞれ記入してください。

① 研究者を志望する動機、目指す研究者像、自己の長所等

② その他、研究者としての資質、研究計画遂行能力を審査員が評価する上で、特に重要と思われる事項（特に優れた学業成績，受賞歴，飛び級入学，留学経験，特色ある学外活動など）

❖ Motivation for being a researcher

I was first introduced to Natural Language processing when I joined a company for internship in the research and development. And my first task was text generation, which led me to my first paper that introduced recurrent neural network (RNN) based language model. This paper proposes an RNN language model applied to speech recognition. Looking at the input and output it only looked like magic and I was fascinated by this magical world of machine learning. I had a chance to apply machine learning to different tasks like optical character recognition and since then I knew that my contribution in making the world a better place to live was going to be through research.

❖ Aim of becoming a researcher

Growing up in a remote village in Africa, I witnessed grinding poverty that was to a large extent the base of crime in the village. This has haunted me and I have always thought of how I could contribute in the fight against such extreme poverty. My aim as a researcher is becoming that person that adds value in transforming the society. In the modern society, a large amount of information is constantly increasing, and this information is valuable to humanity. I therefore would like to make this information valuable in changing the world by discovering valuable information from a huge amount of information and disseminating new techniques to make it of great use. To be able to do this, it is necessary for me to have a diverse experience and interact with other researchers.

❖ Personal Strength

I believe that I posses the following strength:

(i) **In conducting research**, I organize paper-reading sessions with the machine learning Tokyo group, and together we organize monthly hands on workshops to support other young researchers. During the sessions we have a chance to actively discuss with researchers in the industry. Conducting research in various fields, which is important for obtaining a broad perspective as a researcher.

(ii) **Skills and Expertise**, I learnt a lot from the research and development team at Future architect during my internship, I acquired experience in machine learning frameworks, using distributed processing technology, and in the process of working as a research assistant at the Center for Computational sciences of the University of Tsukuba, I have acquired specialized knowledge and technology in natural language process and relational databases using deep learning. Such broad expertise and technology are considered to be important in developing a wide range of research activities.

(iii) **Personal assessment,** I am hands-on and self-motivated individual with a great desire to learn, Self-start, a result oriented person with progressive experience in conducting research, Capable of using teamwork skills to achieve or exceed expectation, Proven track record in managing multiple priorities. During my masters degree at Toyo university I was able to establish a research collaboration with the national university of Rwanda, I am also working on a Project together with Future architect to train a team of AI researchers in Rwanda who will be employed by Future architect to handle the offshore business activity. I also recently started discussions for a possible collaboration between the University of Tsukuba and the University of Rwanda to set up a research center of excellence in artificial intelligence and Robotics at the national University of Rwanda.

[10] T. Mikolov , M. Karafiat, L. Burget, J.Cernock, S.Khudanpur, Recurrent neural network based language model INTERSPEECH 2010