Question Answering from F.A.Q. - Evalita 2016

Armillotta Alessandro - Inversi Alessandro - Savasta Davide

***Abstract*—Questo paper fornisce la soluzione al task QA4FAQ proposto da Evalita. Attraverso la creazione di una rete neurale, data una specifica domanda, si vogliono ottenere le 25 risposte maggiormente correlate. Il progetto realizzato è reperibile al seguente indirizzo GitHub.**

I. Introduction

TROVARE risposte nelle F.A.Q. di una pagina web può risultare un' operazione difficile: gli utenti potrebbero trovare risposte irrilevanti a domande ben specifiche. Le risposte potrebbero trovarsi all'interno della pagina, ma non visualizzabili, perchè la domanda potrebbe essere stata formulata in maniera differente, nel linguaggio naturale.
L'attività proposta permette di rispondere ad una specifica domanda, immessa dall'utente in una query, attraverso la creazione ed elaborazione di una Rete Neurale.
Acquedotto Pugliese (AQP) ha sviluppato un motore di ricerca semantica per le F.A.Q., chiamato AQP Risponde, basato su tecniche di Question Answering (QA). Il sistema consente ai clienti di porre domande in linguaggio naturale, proponendo un elenco di domande frequenti pertinenti e risposte corrispondenti. Inoltre, i clienti possono selezionare una delle domande più frequenti tra quelle recuperate dal sistema e possono fornire il proprio feedback sull'accuratezza percepita della risposta.
Il task QA4FAQ ha l'obiettivo di realizzare una Rete Neurale che permetta di restituire un grado di similarità tra una domanda e tutte le risposte reperibili nel sistema. In base a questa misura di Score, vengono selezionate le top-25.

II. Dataset Description

EVALITA fornisce come training set 406 domande e le relative risposte. In particolare, il dataset è formato da 3 attributi: answer, question e tag. Answer e question vengono utilizzati per addestrare la rete neurale, mentre nella fase di predizione, si inseriscono solo le answer.
All'interno del Test Set sono presenti le 3 domande "Unseen" sulle quali applicare la "prediction". Come descritto sopra la predict consiste nel restituire le 25 risposte più correlate tra le 406 answers disponibili nel dataset originale e la misura di correlazione è rappresentata dallo score di classificazione.

III. System Description

IL SISTEMA è costruito attraverso un framework di apprendimento, il quale prende in ingresso due input: un set di domande (o un set di domande e risposte concatenate) ed un set di risposte e restituisce un indice di similarità tra di esse.
Il processo per l'ottenimento di tale score si articola in alcune fasi fondamentali: Preprocessing (Tokenization, Word2Vec, Pad Sequences) e Modelling. Per poter effettuare le suddette operazioni sono state sfruttate le GPU Tesla P100-PCIE utilizzate in remoto, garantendoci così una potenza di calcolo adeguata alla mole di lavoro prevista. I modelli sono stati realizzati utilizzando le librerie Keras e TensorFlow usato come backend. Keras è una libreria Python minimalista per l'apprendimento approfondito che può essere eseguita su Theano o TensorFlow. È stata sviluppato per rendere veloce e semplice l'implementazione di modelli di deep learning per la ricerca e lo sviluppo.

A. Preprocessing

Nella fase di Preprocessing vengono trattati i dati originali, col fine di trasformarli attraverso tecniche di sostituzione e di tokenizzazione. Inizialmente viene effettuata una procedura di Data Cleaning, eliminando caratteri speciali. Successivamente si passa alla **Tokenizzazione** delle domande e delle risposte: questa fase viene realizzata grazie alle API di "Tanl Italian Pipeline", servizio Web del Dipartimento di Informatica (Pisa) per analizzare testi in lingua italiana. Tokenizzare un testo significa dividere le sequenze di caratteri in unità minime di analisi dette "token":

Il cielo è blu!

["Il", "cielo", "è", "blu", "!"]

Gli elaborati creati vengono memorizzati in diversi file di testo che verranno poi ripresi nelle fasi successive di training e testing delle reti. Effettuata la tokenizzazione viene realizzato il modello **Word2Vec** (si utilizza la libreria Gensim Python) attraverso il quale viene creato il vocabolario delle parole. Word2vec è una rete neurale a 2 layers progettata per elaborare il linguaggio naturale. L'algoritmo richiede in ingresso un corpus (il risultato della tokenizzazione precedente) e restituisce un insieme di vettori che rappresentano la distribuzione semantica delle parole nel testo. Per ogni parola contenuta nel corpus, in modo univoco, viene costruito un vettore in modo da rappresentarla come un punto nello spazio multidimensionale creato. In questo spazio le parole saranno più vicine se riconosciute come semanticamente più simili. Il vocabolario è stato generato utilizzando in particolare i seguenti parametri:

- size = 200, ogni parola nel nostro vocabolario, dopo il training, è rappresentata da un vettore di lunghezza 200
- window = 5 (default), la massima distanza accettata tra la parola corrente e quella predetta all'interno di una frase.

Ogni singola parola è stata poi sostituita con l'indice numerico del vocabolario, generando un vettore come nell'esempio:

["Il", "cielo", "è", "blu", "!"]

[5, 7, 10, 14, 30]

Una volta effettuata la rappresentazione intera, sono state realizzate le corrispondenti **Pad Sequences**, ovvero i vettori di domanda e risposta sono stati resi della stessa dimensione. Ciò è realizzato aggiungendo valori 0 in coda.

Question pre-padding: [5, 7, 10, 14, 30]; len=5
Answer pre-padding:[8, 9, 12, 14, 18, 21, 22, 30]; len=8

Question padded: [5, 7, 10, 14, 30, 0, 0, 0]; len=8
Answer padded: [8, 9, 12, 14, 18, 21, 22, 30]; len=8

Dopo la fase di preprocessing del testo, si è passati alla realizzazione dei modelli.

B. Modelling

Si è proceduto nell'implementare due modelli, al fine di confrontarli ed ottenere la miglior accuratezza. In particolare sono state realizzate due Reti Neurali: il primo modello, **Conv2D - MaxPooling**, sfrutta il prodotto di matrici tra Embedding e Bi-LSTM come input per la rete Convolutional-MaxPooling e un multilayer perceptron (MLP) finale; mentre il secondo modello **AVG Embedding - Bi-LSTM** calcola la media tra le matrici Embedding e Bi-LSTM, successivamente passate ad un MLP. Le reti si basano su processi diversi, ma entrambe sfruttano alcuni dei seguenti layers:

- **Embedding**: L'embedding richiede che i dati in input (stringhe) siano codificati in numeri interi. Questa fase di codifica è stata realizzata durante il preprocessing. In un Embedding, le parole sono rappresentate da vettori densi in cui un vettore rappresenta la proiezione della parola in uno spazio vettoriale continuo. La posizione di una parola all'interno dello spazio vettoriale viene appresa dal testo e si basa sulle parole che circondano la stessa quando viene utilizzata. Il layer prende in input un vettore e per ogni elemento restituisce un vettore di grandezza 200, pari alla rappresentazione dei pesi nel Word2Vec. I pesi vengono utilizzati per misurare l'influenza della parola sullo spazio vettoriale.
- **Bi-LSTM**: L'idea di reti neurali ricorrenti bidirezionali (RNN) è semplice; Implica la duplicazione del layer in input nella rete, in modo tale che ci siano due sequenze affiancate, ottenendo quindi una sequenza originale e una copia invertita della stessa. Questo approccio è stato utilizzato con grande efficacia con le reti neurali ricorrenti a memoria a lungo termine (Long Short-Term Memory). La memoria a lungo termine (LSTM) è un tipo avanzato di rete neurale ricorrente per apprendere le dipendenze a lungo termine all'interno di una sequenza. Quindi, data una sequenza, il Bi-LSTM crea due sequenze LSTM Forward e Backward che apprendono le dipendenze a lungo termine in un senso ed in un altro.
- **Convolutional 2D**: Nel Convolutional dato un kernel/filtro di grandezza k e una finestra MxM, il layer restituisce k mappe di features di grandezza MxM. Il convolutional permette quindi, dati in input dei vettori, l'estrazione di features dagli stessi. Solitamente dopo la convolutional viene applicato il Max Pooling che permette di discretizzare lo spazio vettoriale.
- **Max Pooling**: Il Max Pooling è un processo di discretizzazione basato su campioni. L'obiettivo è di campionare una rappresentazione di input (Matrice), riducendo la sua dimensionalità e consentendo di formulare ipotesi sulle caratteristiche contenute nelle sottoregioni. In pratica permette di estrapolare delle features. Ciò viene fatto in parte per limitare l'overfitting, fornendo una forma astratta della rappresentazione. Inoltre, riduce il costo computazionale riducendo il numero di parametri da apprendere. Dato uno specifico filtro o finestra, il Max Pooling permette di recuperare il valore massimo all'interno di questo filtro, il quale viene fatto scorrere sulla matrice.
- **MLP**: Il Multilayer Perceptron è un modello di rete neurale che mappa insiemi di dati in ingresso in un insieme di dati in uscita appropriati. È fatto di strati multipli di nodi, con ogni strato completamente connesso al successivo. Eccetto che per i nodi d'ingresso, ogni nodo è un neurone con una funzione di attivazione (es.: relu, sigmoid, softmax).

IV. Models

I MODELLI realizzati, come già accennato in precedenza, sono due. Passiamo adesso alla loro analisi.

A. Conv2D-MaxPooling

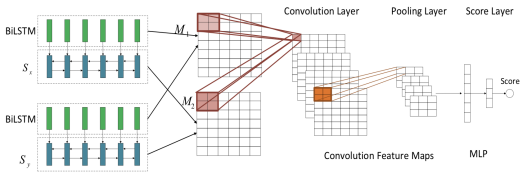

Fig. 1: Convolutional2D - MaxPooling model

Il modello *Conv2D - MaxPooling* (riferimento immagine 9) è formato da 2 layers input, uno per le domande (successivamente domande e risposte concatenate) ed uno per le risposte. Entrambe le sequenze sono rappresentazioni numeriche di parole. I vettori vengono passati ai layers di Embedding, i quali restituiscono un vettore, dove ogni rappresentazione numerica è costituita da 200 valori, i quali a loro volta sono connessi a layers bi-LSTM con un dropout di 0.5. Successivamente, date le trasformazioni vettoriali, viene realizzato un prodotto di matrici. La matrice M1, rappresenta il prodotto delle matrici Embedding di domande (o domande concatenate a risposte) e risposte. La matrice M2, rappresenta il prodotto delle matrici bi-LSTM di domande (o domande concatenate a risposte) e risposte. Le matrici ottenute, M1 e M2, vengono concatenate in un'unica matrice. La matrice risultante costituisce l'input del Convolutional2D. Il convolutional applica ai dati un filtro di grandezza 8, con una finestra di ampiezza 3. Come funzione di attivazione, si è scelta la "relu" ed un successivo dropout pari a 0.5. Il dropout permette di limitare l'overfitting. L'output ottenuto viene passato come input al Max Pooling, il quale

recupera l'elemento massimo in una finestra 2x2. La matrice ottenuta dal Max Pooling è l'input del MLP (200 nodi di input) con funzione di attivazione relu. Si inseriscono altri 100 hidden node con la medesima funzione di attivazione e un ultimo nodo con attivazione sigmoid, il quale fornisce lo score di similarità. La fase di training viene effettuata attraverso la

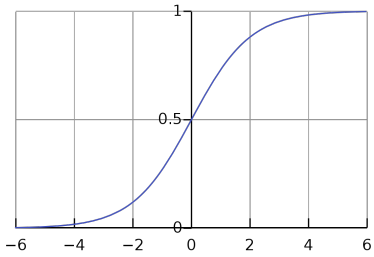

Fig. 2: Sigmoid Function

tecnica dello Stratified K-fold con k=10. Ne risulta che la rete viene addestrata su un train di grandezza 730 ed un test di 82.

B. AVG Embedding - Bi-LSTM

Come per il precedente modello, l'*AVG Embedding - Bi-LSTM*[1] è formato da 2 layers input, uno per le domande (o domande concatenate a risposte) ed uno per le risposte (riferimento immagine 10). Entrambe le sequenze sono rappresentazioni numeriche di parole. I vettori vengono passati ai layers Embedding che restituiscono un vettore dove ogni rappresentazione numerica è rappresentata da 200 valori, i quali a loro volta sono connessi a layers bi-LSTM con un dropout di 0.5. A differenza del precedente modello, si calcola la media dei vettori ottenuti dall'Embedding e dal Bi-LSTM e successivamente si concatenano i valori. La concatenazione viene passata direttamente al MLP e non al Conv - Max-Pooling. Il MLP ha, quindi, 200 nodi di input con funzione di attivazione relu, altri 100 hidden node con funzione di attivazione relu e un ultimo nodo con attivazione sigmoid, il quale fornisce lo score di similarità. Anche in questo caso, per la fase di training, è stato utilizzato lo Stratified K-Fold con k=10. Anche in questo caso la rete viene allenata su un train di grandezza 730 ed un test di 82.

V. Training & Results

Dopo aver descritto i modelli utilizzati per il calcolo della similarità, procediamo con l'analisi dei risultati modello per modello.
La fase di Training è stata eseguita in due modalità di input:

1) **Question - Answer**: entrambi i modelli prendono in input le Question e le Answer;
2) **Question - Question + Answer**: in questo modo, i modelli prendono in input le Question da una parte e dall'altra un vettore contenente question e Answer.

Le migliori **epoch** sono state salvate attraverso la funzione *ModelCheckpoint*, la quale fa parte di una vasta serie di funzioni *"Callbacks"*. La funzione, permette di salvare il modello in base al miglior valore di misura considerato (accuracy o loss). Nei modelli creati, avendo avuto molte problematiche computazionali, si è scelto di prendere in considerazione il loss della validation. Inoltre, è stata utilizzata la funzione *EarlyStopping* per prevenire l'overfitting, oltre all'utilizzo di diversi Dropout. Infine è stata utilizzata la funzione *CSVLogger* per esportare i risultati su un CSV.
Il dataset utilizzato per la fase di training era formato da 406 domande e 406 risposte. Per aumentare le performance della rete, è stato raddoppiato il dataset con 406 question-answer corrette (class label = 1) e 406 question-answer errate scelte random (class label = 0).
Di seguito, nella tabella, vengono mostrati i migliori risultati ottenuti utilizzando il dataset Q-A e Q-QA in entrambi i modelli:

Conv2D - MaxPooling		
	Train Accurancy	Test Accurancy
Q-A	99%	100%
Q-QA	100%	100%
AVG Embedding - Bi-LSTM		
	Train Accurancy	Test Accurancy
Q-A	49%	49%
Q-QA	52%	50%

TABLE I: Train and Test Accuracy

Come si può ben notare dalla tabella, i risultati migliori sono stati ottenuti sul modello Conv2D - MaxPooling. Sono state portare a termine svariate prove, aumentando o diminuendo batch, value stop, epoch e dropout, ma i risultati del modello AVG Embedding - Bi-LSTM non sono cambiati. Le accuratezze, sia di train che test, oscillano tra 55 e 30, senza alcun minimo segno di miglioramento.

A. Conv2D - MaxPooling (Q-A) Results

Il modello *Conv2D - MaxPooling (Q-A)* è stato addestrato con uno *Stratified k-fold* con k=10. Ad ogni K è stato impostato un batch_size=12 e epochs=50. Per il training sono stati utilizzati 730 records e 82 per il test.

L'epoch selezionata permette di ottenere un'accuratezza sul test pari al 100% che, visto gli scarsi risultati ottenuti senza applicare lo Stratified k-fold, sembra essere ottimale e impossibile da migliorare. L'accuratezza così alta è stata ottenuta sicuramente per il grande numero di K, ma qualsiasi altro valore al di sotto di K=10 restituiva valori tra 50% e 40%.
Dopo aver analizzato l'accuratezza, è stato effettuata la *Predict* sulle domande di test. Il task chiede di restituire le 25 risposte più correlate alla domanda (alta similarità). Per problemi di spazio, visto la lunghezza delle risposte, riportiamo solo alcune delle 25 più correlate.

1) ***Come posso pagare la bolletta?****:*

- Per i pagamenti dall' estero delle fatture di Acquedotto Pugliese , e possibile utilizzare i servizi di Home Banking riportando il numero di MAV come presente sul bollettino allegato alla fattura Inoltre , solo qualora non si potesse pagare con altri metodi , di seguito si riportano le

[1]Riferimento immagini pag 6.

coordinate IBAN : n - IBAN : IT27 O076 0104 0000 0001 6240 731 (ove il quinto carattere e rappresentato dalla lettera O di Omega , e coincide con il codice CIN) n - BIC / SWIFT : BPPIITRRXXX... **(similarity 100%))**

- Il pagamento potra essere effettuato con bonifico bancario. Per i pagamenti dall' estero delle fatture di Acquedotto Pugliese , e possibile utilizzare i servizi di Home Banking riportando il numero di MAV come presente sul bollettino allegato alla fattura Inoltre , solo qualora non si potesse pagare con altri metodi , di seguito si riportano le coordinate IBAN : n - IBAN : IT27 O076 0104 0000 0001 6240 731 (ove il quinto carattere e rappresentato dalla lettera O di Omega , e coincide con il codice CIN) n - BIC / SWIFT : BPPIITRRXXX... **(similarity 100%)**
- Utilizzando i bollettini MAV le fatture possono essere pagate presso : uno degli uffici postali dislocati su tutto il territorio nazionale : qualsiasi sportello bancario aderente al servizio MAV , in maniera gratuita ; Tabaccheria o Ricevitoria del Lotto PuntoLis o Ricevitorie Sisal dislocate su tutto il territorio nazionale... **(similarity 99%)**

2) ***Cos'è il deposito cauzionale?****:*

- Il deposito cauzionale e una somma di denaro che l' utente versa al gestore a titolo di garanzia e che deve essere restituita dopo la cessazione del contratto nel rispetto delle condizioni contrattuali in vigore. **(similarity 99%)**
- Il deposito cauzionale e applicato a tutti i clienti Sono esclusi dall' addebito : le utenze ad uso pubblico ; gli utenti beneficiari di bonus idrico ; le utenze con domiciliazione bancaria / postale che abbiano avuto consumi inferiori a 500 mc nell' anno solare precedente Sono previste forme di garanzia alternative al deposito cauzionale per gli usi diversi dal domestico con consumi superiori a 500 mc / anno **(similarity 99%)**
- Dal 01.06.2014 è entrato in vigore il metodo di calcolo del deposito cauzionale secondo le disposizioni indicate da AEEGSI con delibere n.86 del 28.02.2013 e n.643 del 27.12.2013 Tali disposizioni prevedono che il calcolo del deposito cauzionale sia applicato ad ogni utente , in base al corrispettivo trimestrale , delle quote fisse e variabili... **(similarity 99%)**

3) ***Ho bisogno di aiuto.****:*

- Qualora il contratto di somministrazione del locale ceduto in affitto sia intestato al locatario , a fine locazione , il titolare del contratto dovrà dare comunicazione scritta controfirmata all' indirizzo dell' ufficio assistenza clienti territorialmente competente allegando il documento di riconoscimento e ultima fattura , più recapito telefonico ; la domanda potrà essere inoltrata anche via pec all' indirizzo clienti@pec.aqp.it... **(similarity 100%)**
- Qualora si sia già inoltrata richiesta di allacciamento e si fosse già concordato un appuntamento , in fase di sopralluogo il tecnico-commerciale , al fine del completamento del contratto ad uso occasionale e provvisorio , necessita dei seguenti documenti in relazione con la forma giuridica dell' intestatario del servizio.. **(similarity 100%)**
- Gli uffici commerciali AQP sono aperti il lunedi , mercoledi e venerdi dalle ore 9.00 alle ore 13.00 ; il martedi e il giovedi dalle ore 9.00 alle ore 13.00 e dalle ore 15.00 alle ore 17.00 Per verificare puntualmente gli indirizzi e orari dei singoli uf.fici commerciali consulti il sito www.aqp.it.. **(similarity 99%)**

B. Conv2D - MaxPooling (Q - Q+A) results

1) ***Come posso pagare la bolletta?****:*

- I moduli contrattuali , in particolare per l' uso domestico , sono attribuiti in relazione con il numero di unità abitative di cui l' immobile servito si compone , come rilevabile da documento catastale e / o altro documento dell' immobile Pertanto , qualora nel corso del tempo , con sopraggiunti cambiamenti **(similarity 100%)**
- Tutti i riferimenti e gli indirizzi e / o i canali di contatto con Acquedotto Pugliese sono disponibili sul portale internet www.aqp.it , nella sezione contatti.**(similarity 100%)**
- Tutti i clienti che abbiano un regolare contratto di somministrazione idrica e / o integrata con Acquedotto Pugliese sono tenuti ad effettuare il pagamento delle fatture , come previsto dalle condizioni generali della fornitura del servizio Il corrispettivo del servizio fatturato viene determinato sulla base dei consumi registrati dal misuratore Inoltre , qualora l' utente prelevi acqua da fonti alternative o abbia un approvvigionamento autonomo , secondo le norme vigenti , è vietata qualsiasi connessione tra gli impianti interni diversamente alimentati**(similarity 100%)**

Come si può notare dai risultati, l'accuratezza risulta essere massima, ma le risposte non sono correlate.

2) ***Cos'è il deposito cauzionale?****:*

- Nel caso in cui l' utente destinatario del Bonus Idrico non ricevesse comunicazione scritta , potrà richiedere copia della stessa direttamente presso i nostri uffici territorialmente competenti o contattando il numero verde commerciale 800.085.853 In entrambi i casi deve fornire il numero di protocollo della domanda ed il codice fiscale. **(similarity 100%)**
- Il numero delle utenze servite da Acquedotto Pugliese è di oltre 970.000 Le utenze allacciate garantiscono l' accesso all' acqua ad oltre 4 milioni di abitanti della Puglia e della provincia di Avellino **(similarity 100%)**
- Le fatture di Acquedotto Pugliese , sia per quanto attiene il layout che i contenuti , seguono le indicazioni dell' Autorità per l' Energia Elettrica il Gas e il Servizio Idrico (AEEGSI) , in conformità con la Direttiva per la trasparenza dei documenti d **(similarity 100%)**

Anche in questo caso l'accuratezza è massima, ma le risposte risultano essere, ancora una volta, non correlate.

3) ***Ho bisogno di aiuto.****:*

- È possibile inoltrare reclamo per inversione contatori attraverso i seguenti canali : di persona presso gli Uffici commerciali o gli Sportelli comunali on line ; telefonando al numero verde 800.085.853 ; **(similarity 100%)**
- Il modulo o impegnativa contrattuale e attribuita per un contratto di utenza a uso domestico in base al numero

di unita abitative , pertanto , se dovessero aumentare le unita immobiliari servite dallo stesso contatore , e necessario**(similarity 100%)**
- L' acqua è presente in natura e sulla terra gli oceani coprono più del 70% della superficie terrestre e contengono una quantita di 1.350 milioni di chilometri **(similarity 100%)**

Per quest'ultima domanda, le risposte sembrerebbero correlate alla domanda, ma i risultati ottenuti per le precedenti non ci permettono di garantire effettivamente l'affidabilità del modello in questione.

C. AVG Embedding - Bi-LSTM (Q - A) results

Il modello AVG Embedding - Bi-LSTM è stato addestrato con uno Stratified k-fold con k=10 e ad ogni Ki è stato impostato un batch_size=10 e epochs=50. Per il training sono stati utilizzati 730 records e 82 per il test.

L'epoch selezionata permette di ottenere un'accuratezza sul test pari al 48% che, visto gli scarsi risultati ottenuti senza applicare lo Stratified k-fold, sembra essere ottimale e quasi impossibile da migliorare. L'accuratezza sul training invece non ha mai superato il valore di 52%. Dopo aver analizzato l'accuratezza, è stato effettuata la predict sulle domande di test. Il task chiede di restituire le 25 risposte più correlate alla domanda. Per problemi di spazio, visto la lunghezza delle risposte, riportiamo solo alcune delle 25 più correlate.

1) ***Come posso pagare la bolletta?****:*
- In questa circostanza si suggerisce di effettuare la domiciliazione bancaria o postale delle fatture con addebito automatico SEPA in conto corrente Per aderire alla domiciliazione con addebito automatico in conto è sufficiente scaricare il modulo disponibile su www.aqp.it nella sezione clienti / fattura / modalità di pagamento ... **(similarity 56%)**
- Il modello per la dichiarazione dei dati catastali , della concessione edilizia e di altre informazioni necessarie sull' insediamento da servire è consegnato , ed è preferibile che sia compilato e sottoscritto , in fase di sopralluogo da parte del tecnico AQP , è denominato DICHIARAZIONE SOSTITUTIVA DI CERTIFICAZIONE E DELL' ATTO DI NOTORIETA' (ex Artt 46 e 47 D.P.R 445/2000 e successive modifiche e integrazioni) **(similarity 54%)**
- Quando si richiede un nuovo allacciamento distaccandosi da un impianto esistente (ad esempio nel caso del distacco di una unità abitativa da un condominio) , il titolare dell' impianto esistente deve richiedere una innovazione contrattuale per ridurre il numero di moduli contrattuali dell' unità distaccata In tale circostanza , nella fattura di chiusura sarà restituito tutto il deposito cauzionale dell' impianto esistente e nella prima fattura del contratto con un modulo in meno sarà addebitato il nuovo deposito adeguato **(similarity 50%)**

2) ***Cos'è il deposito cauzionale?****:*
- In questa circostanza si suggerisce di effettuare la domiciliazione bancaria o postale delle fatture con addebito automatico SEPA in conto corrente Per aderire alla domiciliazione con addebito automatico in conto è sufficiente scaricare il modulo disponibile su www.aqp.it nella sezione clienti / fattura / modalità di pagamento...**(similarity 74%)**
- Il modello per la dichiarazione dei dati catastali , della concessione edilizia e di altre informazioni necessarie sull' insediamento da servire è consegnato , ed è preferibile che sia compilato e sottoscritto , in fase di sopralluogo da parte del tecnico AQP , è denominato DICHIARAZIONE SOSTITUTIVA DI CERTIFICAZIONE E DELL' ATTO DI NOTORIETA' (ex Artt 46 e 47 D.P.R 445/2000 e successive modifiche e integrazioni).**(similarity 54%)**
- Quando si richiede un nuovo allacciamento distaccandosi da un impianto esistente (ad esempio nel caso del distacco di una unità abitativa da un condominio) , il titolare dell' impianto esistente deve richiedere una innovazione contrattuale per ridurre il numero di moduli contrattuali dell' unità distaccata In tale circostanza , nella fattura di chiusura sarà restituito tutto il deposito cauzionale dell' impianto esistente e nella prima fattura del contratto con un modulo in meno sarà addebitato il nuovo deposito adeguato.**(similarity 50%)**

3) ***Ho bisogno di aiuto.****:*
- In questa circostanza si suggerisce di effettuare la domiciliazione bancaria o postale delle fatture con addebito automatico SEPA in conto corrente Per aderire alla domiciliazione con addebito automatico in conto è sufficiente scaricare il modulo disponibile su www.aqp.it nella sezione clienti / fattura / modalità di pagamento...**(similarity 74%)**
- Il modello per la dichiarazione dei dati catastali , della concessione edilizia e di altre informazioni necessarie sull' insediamento da servire è consegnato , ed è preferibile che sia compilato e sottoscritto , in fase di sopralluogo da parte del tecnico AQP , è denominato DICHIARAZIONE SOSTITUTIVA DI CERTIFICAZIONE E DELL' ATTO DI NOTORIETA' (ex Artt 46 e 47 D.P.R 445/2000 e successive modifiche e integrazioni)**(similarity 53%)**
- Quando si richiede un nuovo allacciamento distaccandosi da un impianto esistente (ad esempio nel caso del distacco di una unità abitativa da un condominio) , il titolare dell' impianto esistente deve richiedere una innovazione contrattuale per ridurre il numero di moduli contrattuali dell' unità distaccata In tale circostanza , nella fattura di chiusura sarà restituito tutto il deposito cauzionale dell' impianto esistente e nella prima fattura del contratto con un modulo in meno sarà addebitato il nuovo deposito adeguato. **(similarity 52%)**

D. AVG Embedding - Bi-LSTM (Q - QA) results

Per quanto riguardo il training del modello AVG Embedding - Bi-LSTM (Q - QA), dati gli scarsi risultati ottenuti nel precedente, è stato deciso di non riportare i risultati. Nel caso si volessere visionare i risultati, si rimanda alla repository di GitHub nella cartella "result".

VI. Experiment

Per testare l'affidabilità del modello prodotto, verifichiamo il suo funzionamento inserendo altre tre domande, non presenti nel dataset originale.

- **Qual è il costo dell'acqua** :
  A parte i costi energetici, in caso di utilizzo di un sistema di autoclave, il costo di 1 litro di acqua al rubinetto e di Euro 0,001, a tariffa agevolata per uso domestico, inclusi i costi di allontanamento e depurazione... (**similarity 94%**)
- **Ho rilevato un consumo anomalo** :
  Per segnalare un guasto o in caso di una emergenza riguardante il servizio idrico, fognante o depurativo nell'ambito servito da Acquedotto Pugliese comporre il numero verde 800.735.735 operativo tutti i giorni 24... (**similarity 98%**)
- **Come posso richiedere il bonus idrico** :
  Nel caso in cui il cittadino fosse intestatario della fornitura idrica, ricevera direttamente nella fattura consumi AQP l'accredito del Bonus Idrico in un'unica soluzione, se dovuto.... (**similarity 99%**)

Queste sperimentazioni sono state svolte utilizzando il Conv2D - MaxPooling (Q-A) poichè ha manifestato delle prestazioni nettamente superiori rispetto all'addestramento con Question-QuestionAnswer e al modello AVG Embedding - Bi-LSTM.

VII. Evaluation

Si procede alla fase di validazione del modello dato il Test Set completo fornito da Evalita. Per ogni domanda nel dataset (1132 domande), si devono restituire 25 possibili risposte, ordinate secondo uno score scelto dai partecipanti. I vari sistemi sono stati classificati in base all'**accuracy@1(c@1)**. Evalita calcola la precisione del sistema, prendendo in input solamente la prima risposta corretta. Questa metrica è usata per il ranking finale. La formula di c@1 è:

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n})$$

Fig. 3: accuracy@1(c@1)

dove nR numero di domande correttamente risposte, nU il numero di domande non risposte (0), e n il numero totale di domande.

Applicando la formula sul nostro miglior modello (Conv2D-MaxPooling QA) il risultato ottenuto è il seguente:

0.1657

Confrontando l'overview di Evalita 2016, la baseline presenta una C@1 pari a 0.4076, perciò possiamo affermare che il nostro modello presenta delle performance scarse.
Per questo motivo abbiamo deciso di effettuare altri esperimenti, partendo dal riesaminare l'embedding.

VIII. Embedding Experiments

A partire dal nostro modello base (Conv2D - MaxPooling) sono stati effettuati vari esperimenti utilizzando come pesi dell'embedding un W2V pretrained. In particolare è stato inserito un W2V estratto da Wikipedia Italian, fornitoci con vettore di dimensione 50. Il modello base è stato modificato sostituendo i vettori di grandezza 200 (ogni parola rappresentata da un vettore di lunghezza 200) con vettori di grandezza 50 (riferimento immagine 9).

- Come primo esperimento sono state utilizzate le matrice dei pesi sia nell'embedding input A che nell'embedding input Q. In fase di training i pesi non venivano aggiornati e i valori di accuratezza rimanevano invariati per tutte le epoche (0.0).
- Come secondo esperimento abbiamo effettuato una copia della matrice dei pesi per evitare eventuali conflitti, cercando così di migliorare le performance. Anche in questo caso i risultati non sono stati soddisfacenti presentando valori di accuratezza invariati (0.0)
- Come terzo caso abbiamo generato un unico layer di embedding (riferimento immagine 11) sia per A che per Q. Anche in questo caso abbiamo riscontrato le medesime problematiche precedenti.
- Come ultimo esperimento abbiamo utilizzato la matrice dei pesi solo per l'embedding input A, ottenendo un netto miglioramento delle performance in termini di accuratezza. In questo caso i pesi venivano aggiornati ad ogni epoca.

Analizzando le varie prove, abbiamo riscontrato una piena funzionalità solo nel caso in cui la matrice dei pesi veniva inserita in un singolo layer. Ciò ci fa dedurre che esista un possibile bug in Keras, qualora presente in entrambi. Di seguito riportiamo le performance (accuracy, c@1) dei nuovi modelli.

IX. Evaluation Results

Per i nuovi modelli abbiamo utilizzato un batch_size pari a 25 con 800 epoche. Di seguito vengono riportati i risultati.

Conv2D - MaxPooling			
	Train Accurancy	Test Accurancy	C@1
Q-A (50)	99%	50%	0.4389
Q-QA (50)	99%	43%	0.2673
Q-A (200)	99%	100%	0.1657

TABLE II: Train and Test Accuracy. C@1

Osservando i risultati in tabella, il miglior modello in termini di C@1 risulta essere il Conv2D-MaxPooling QA con vettori di dimensione 50, restituendo un valore maggiore rispetto alla baseline di Evalita 2016. I plot relativi alle suddette metriche sono riportati a pagina 11

X. Conclusion

Comparando i risultati derivanti dall'applicazione dei due diversi modelli, possiamo subito notare una netta differenza tra il Conv2D - MaxPooling model e l'AVG Embedding - Bi-LSTM model. Tale scarto tra i valori di accuratezza deriva

dall'utilizzo del **prodotto di matrici** e della rete **Convolutional 2D** (nel primo modello), grazie ai quali siamo in grado di aumentare la precisione della nostra analisi.
Sempre dai risultati, si può notare che una rete allenata su Embedding pesati, produce risultati migliori. Sarebbe utile poter valutare le reti create, potendo pesare entrambi gli Embedding, senza alcun bug. In futuro, sarebbe utili effettuare altri esperimenti su questi modelli. Per esempio, si potrebbe provare a pesare l'Embedding input Q invece di A. Allo stato attuale, usando l'C@1 come valore di rank, il nostro ultimo modello Conv2D-MaxPooling, risulta essere il più performante e superiore alla baseline di Evalita.

REFERENCES

[1] Wenzheng Feng, Yu Wu, Wei Wu, Zhoujun Li, Ming Zhou. Beihang-MSRA at SemEval-2017 Task 3: A Ranking System with Neural Matching Features for Community Question Answering. https://www.aclweb.org/anthology/S/S17/S17-2045.pdf.
[2] Keras Documentation. https://keras.io.
[3] TensorFlow Documentation. https://www.tensorflow.org.
[4] QA4FAQ @ EVALITA 2016 - Question Answering for Frequently Asked Questions Task. http://qa4faq.github.io.
[5] The Italian Tanl Server Web service. http://tanl.di.unipi.it/it/api.

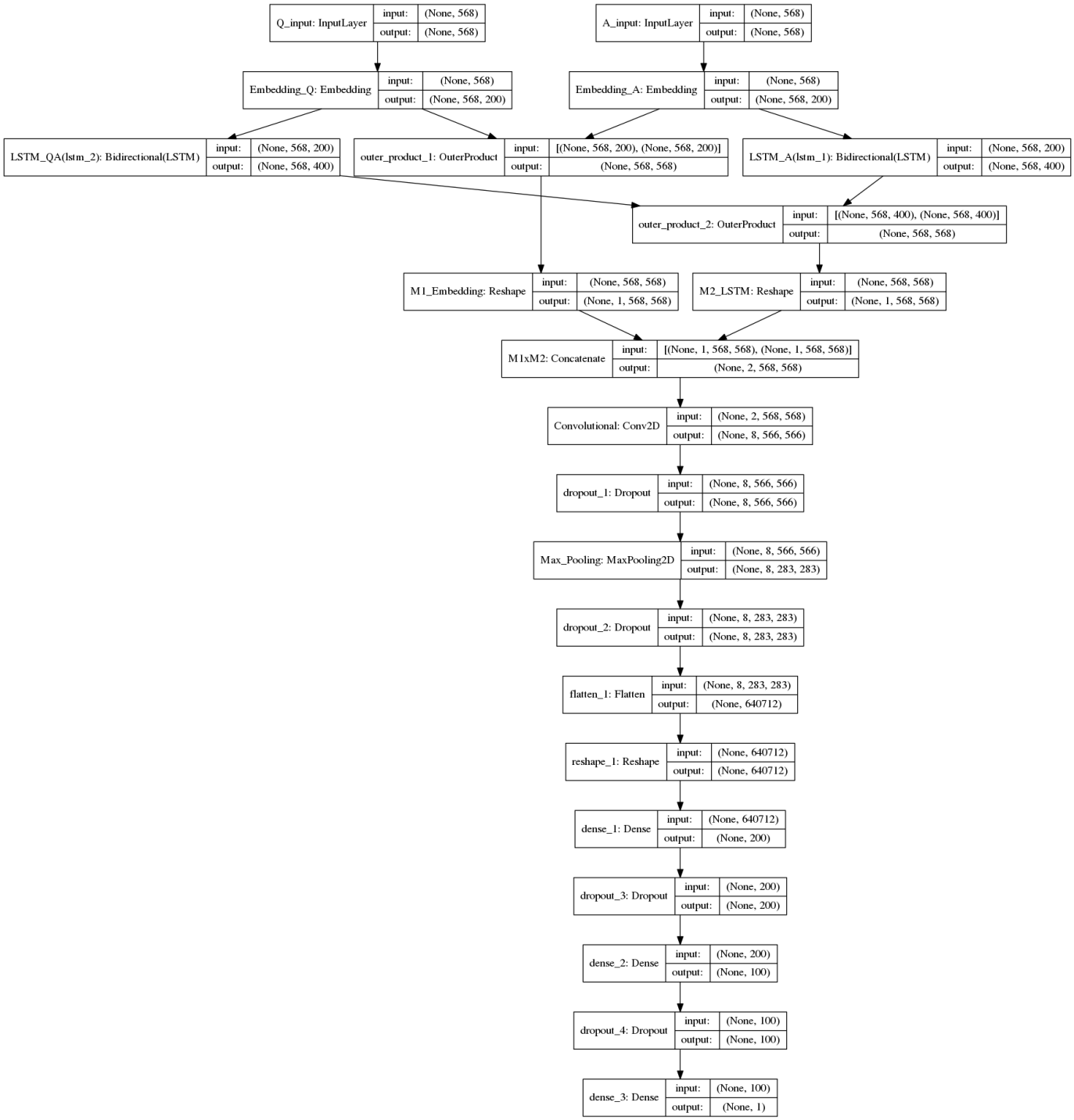

Fig. 4: Conv2D - MaxPooling with Embedding 200

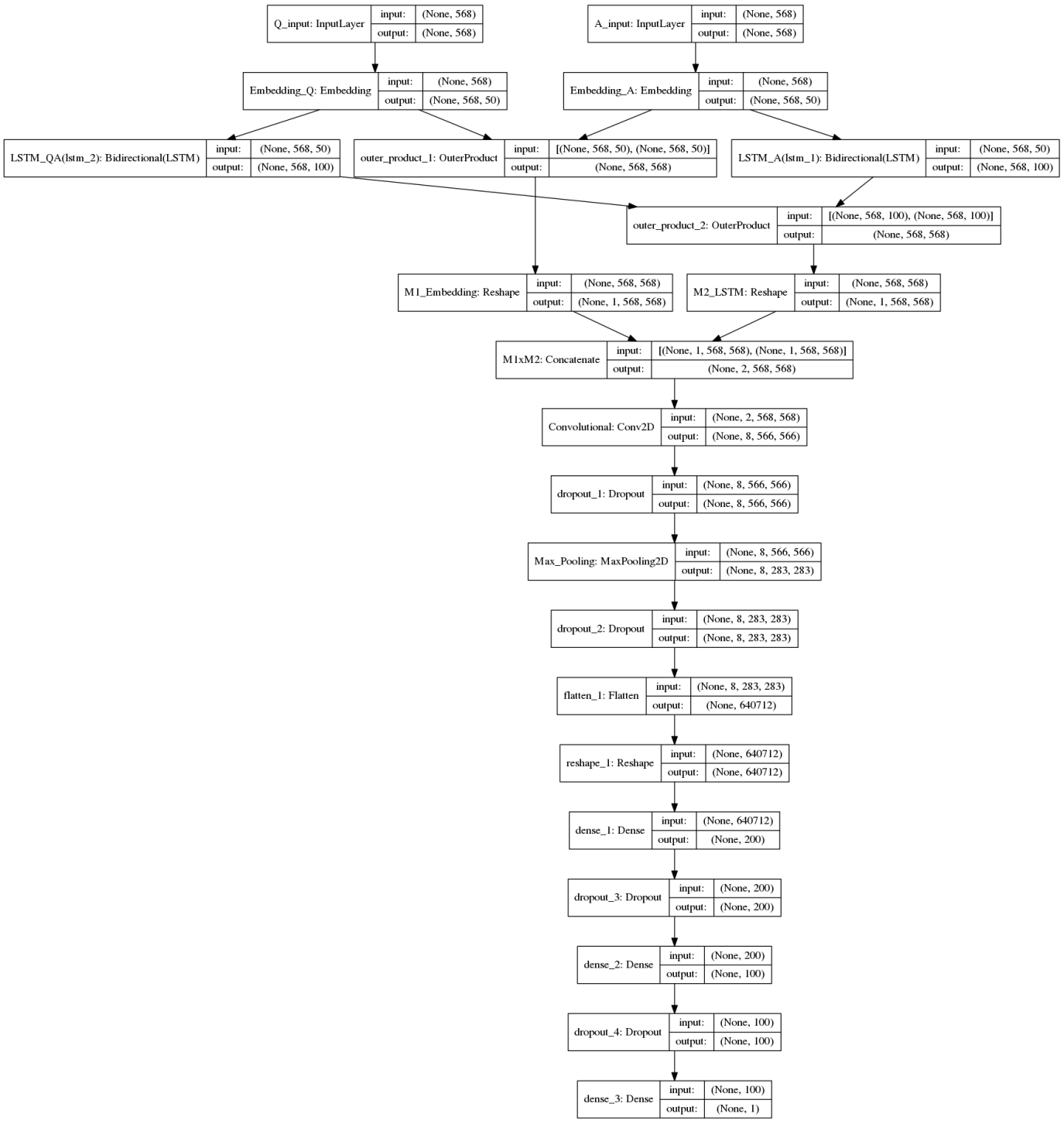

Fig. 5: Conv2D - MaxPooling with Embedding 50

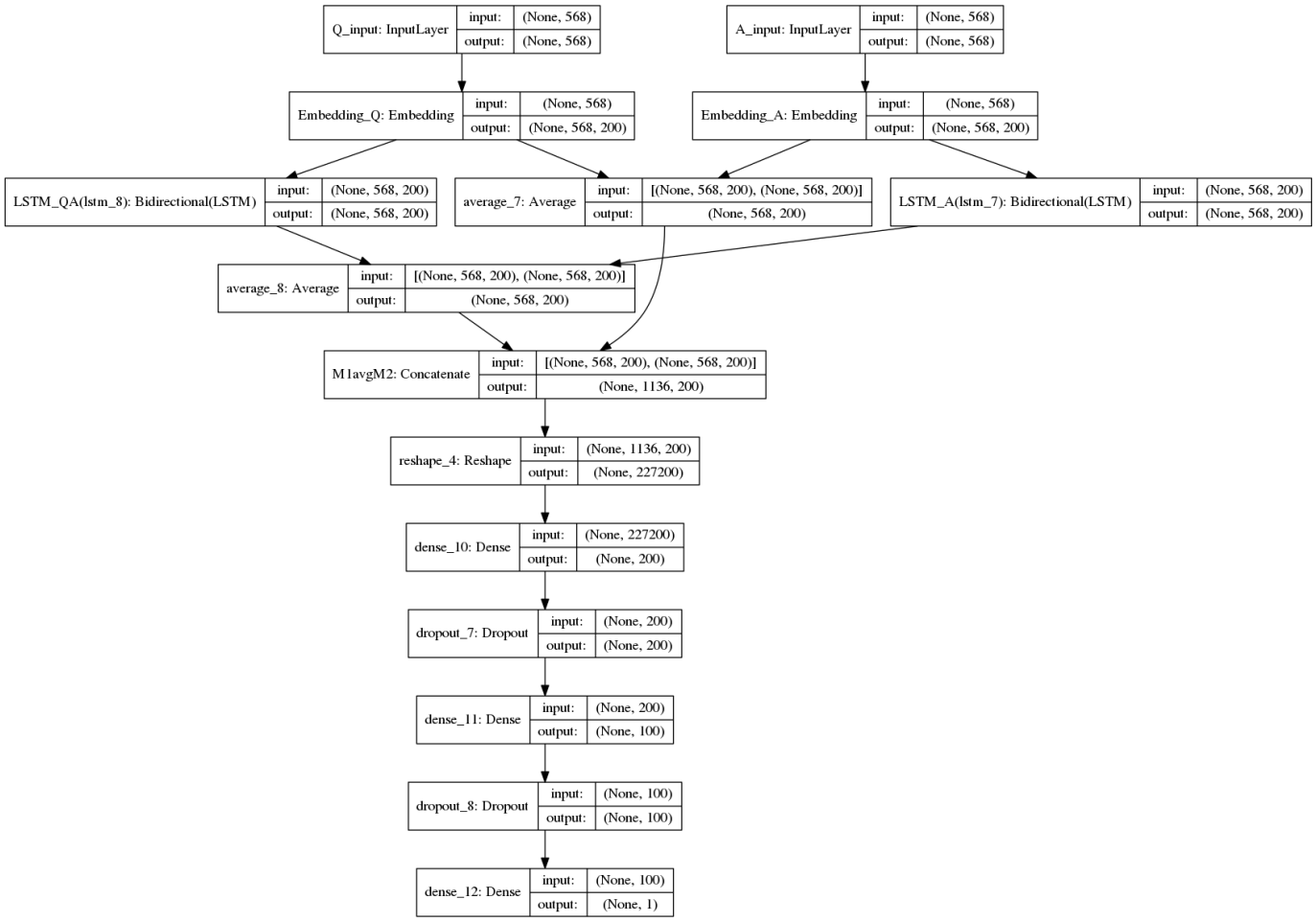

Fig. 6: Avg Embedding - Bi-LSTM

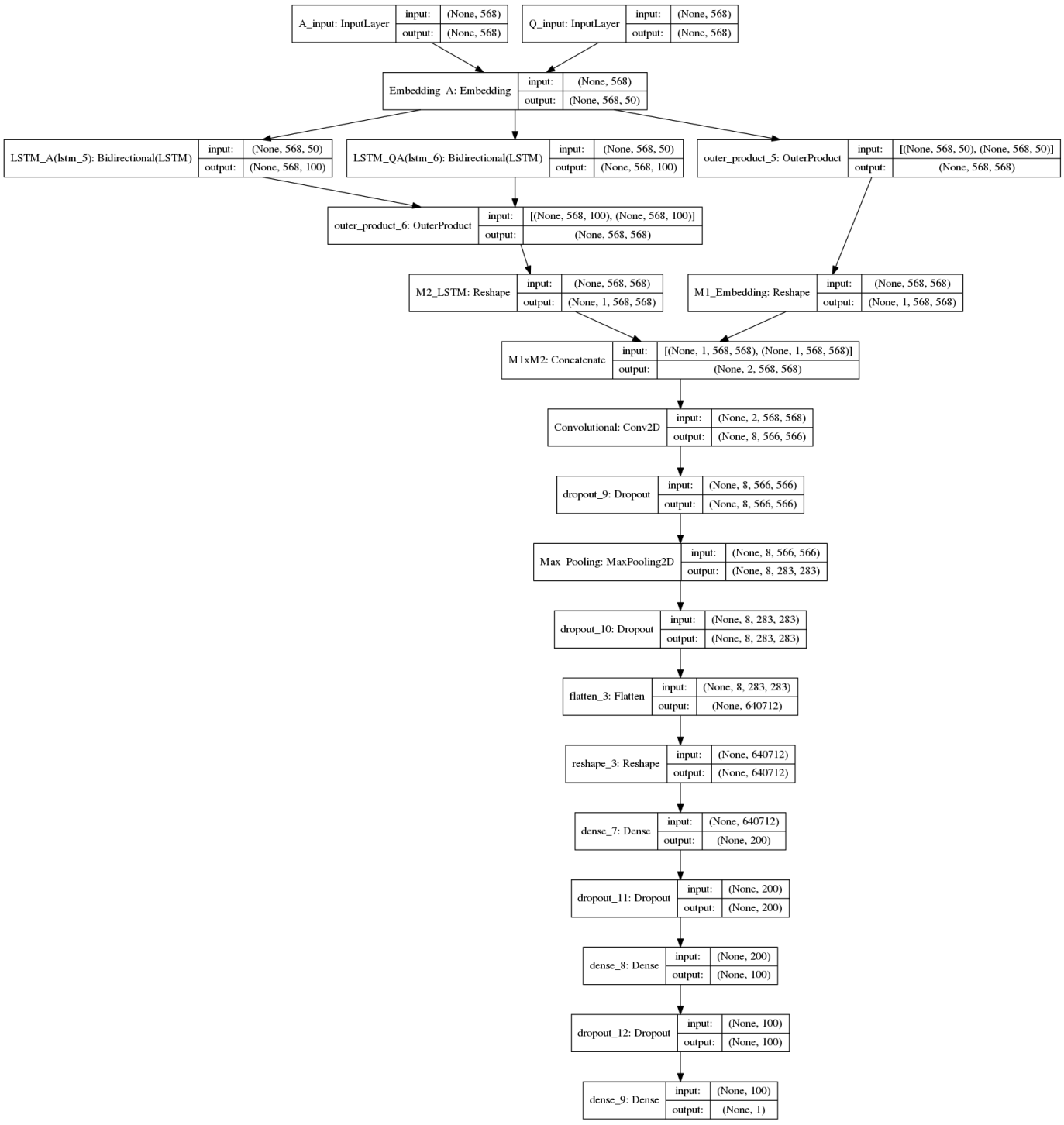

Fig. 7: Conv2D - MaxPooling with one Embedding Layer

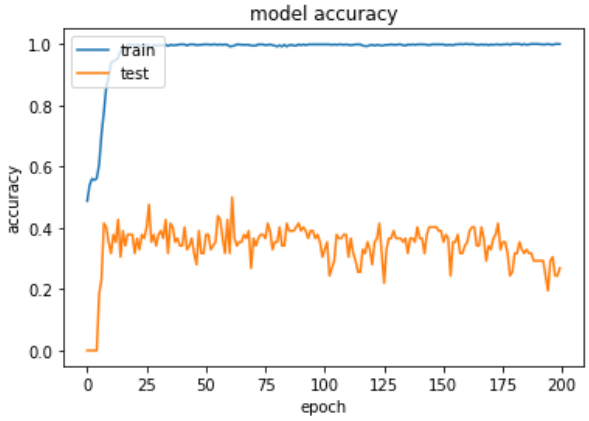

Fig. 8: Accuracy QA

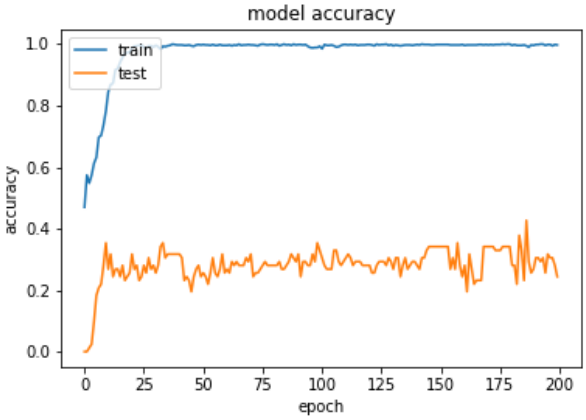

Fig. 11: Accuracy QQA

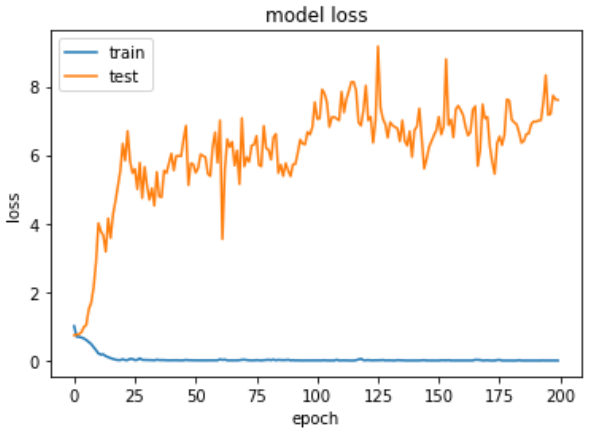

Fig. 9: Loss QA

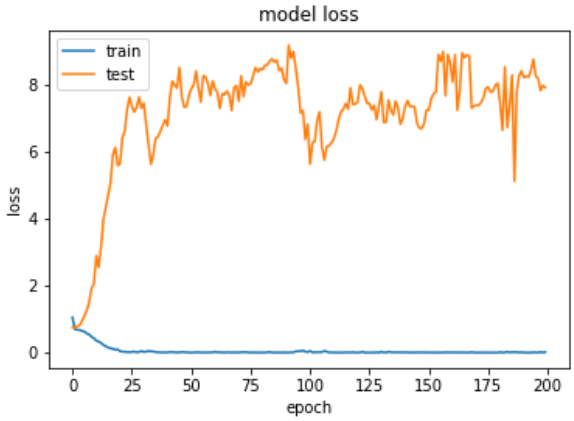

Fig. 12: Loss QQA

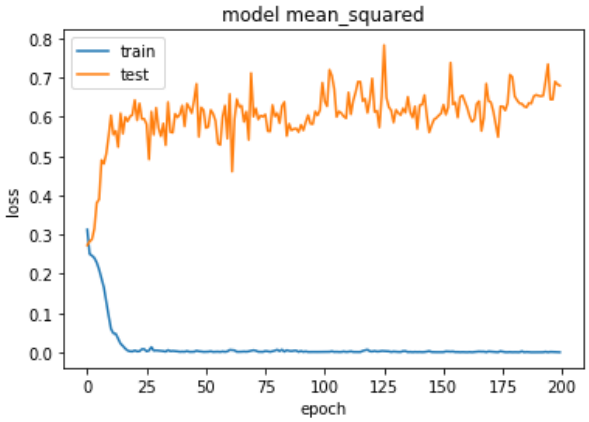

Fig. 10: Mean Squared QA

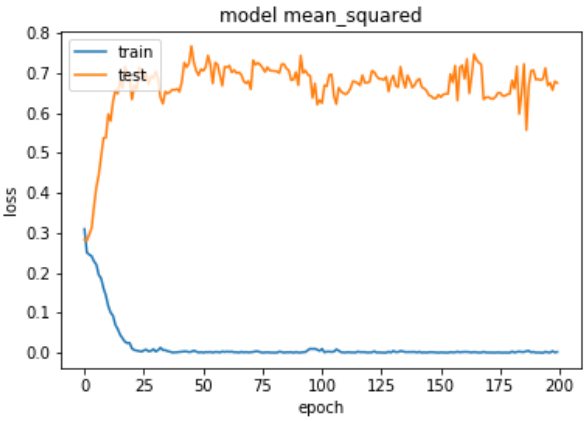

Fig. 13: Mean Squared QQA