Isabox: Question Answering interattivo per le basi di conoscenza dei Comuni italiani

Philippe Binet[2], Marco Bressan[1], Simone Cangialosi[2], Davide Simionato[1], Guido Vetere[2]

[1]Boxxapps S.r.l., [2]Isagog S.r.l.

p.binet@isagog.com, marco.bressan@boxxapps.com, s.cangialosi@isagog.com,
davide.simionato@boxxapps.com, g.vetere@isagog.com

1 Introduzione

l'Italia ha circa 8000 Amministrazioni locali che, con l'espandersi delle tecnologie dell'informazione, si trovano oggi ad affrontare complessi problemi di gestione del patrimonio informativo. Le informazioni amministrate sono infatti sempre più complesse, strutturate e raccolte in svariate forme, molteplici ed eterogenei formati, contenitori, supporti e luoghi, e tuttavia la loro accessibilità deve essere garantita. Accade spesso invece che tali informazioni siano di fatto inattingibili e l'utente si trovi ostaggio di inconsapevoli "rapitori" che (loro malgrado) sono gli unici a detenere la chiave per il loro accesso. Ed è proprio dai cittadini e dagli amministratori che nasce la richiesta di poter accedere alle informazioni in modo intuitivo, naturale, indipendente, interamente svincolato dalla conoscenza dello strumento informatico, potendosi così concentrare sulle necessità e sull'oggetto della ricerca.

L'Intelligenza Artificiale (IA) è vista oggi come un elemento fondamentale della strategia per il miglioramento dei servizi della Pubblica Amministrazione[1]. In particolare, il trattamento del linguaggio naturale (NLP) sperimentato nei sistemi di conversazione ed interrogazione (QA) e le metodologie di rappresentazione della conoscenza (KR) offrono, nei loro recenti sviluppi e mediante opportune integrazioni, la possibilità di offrire soluzioni efficaci alle esigenze informative delle Amministrazioni.

Il progetto Isabox nasce dalla collaborazione di Boxxapps (System Integrator per la Pubblica Amministrazione) e Isagog (Startup innovativa di Intelligenza Artificiale) ed ha per scopo lo sviluppo di una piattaforma per l'interrogazione interattiva del patrimonio informativo comunale. Il presente contributo offre una sintesi del disegno del sistema, delle metodologie e delle assunzioni che vi sono alla base, e degli aspetti innovativi della soluzione.

2 Interactive Question Answering

Per Question Answering interattivo (IQA nel seguito) si intende l'interrogazione di una sorgente di dati attraverso un'interazione con l'utente che ha lo scopo di disambiguare, precisare o arricchire la formulazione della *query* al fine di ottenere risposte specifiche e soddisfacenti. La ricerca in questo campo è attiva da molte decadi ed è stata condotta, attraverso le varie stagioni dell'IA, con approcci e tecniche di varia natura [Webb e Webber, 2009]. I sistemi di IQA si possono caratterizzare sia come sistemi di dialogo specializzati per il task del Question Aswering, sia, per converso, come sistemi di interrogazione di sorgenti informative in linguaggio naturale dotati di capacità di dialogo [Soares e Parreiras, 2018].

L'assunzione che l'utente abbia lo scopo specifico di reperire informazione permette di caratterizzare l'interazione come *system driven* e di trattare alcuni fenomeni tipici delle conversazioni naturali, come ad esempio il *topic shift* [Fernández, 2014]. D'altro canto, l'interazione con l'utente permette in linea di principio di superare i problemi di espressività linguistica di cui soffrono i sistemi di information retrieval, nei quali l'incompletezza o la vaghezza delle domande limitano spesso le prestazioni informative. L'ipotesi progettuale è che tali presupposti facilitino la realizzazione di piattaforme efficaci, robuste e di facile gestione, abilitando soluzioni industrialmente mature.

Nonostante la relativa semplicità dei sistemi di IQA nel quadro generale dei sistemi conversazionali e dei sistemi di interrogazione in linguaggio naturale, molti dei problemi da risolvere non sono banali e sono tutt'oggi oggetto di ricerca. Rispetto ad altri sistemi di question answering su sorgenti strutturate sperimentati in passato (ad esempio QALLME [2] [Magnini *et al.*, 2009]), una delle principali caratteristiche di Isabox consiste nell'integrazione di avanzate tecniche di analisi linguistica (dipendenze sintattico-concettuali) con le recenti metodologie di mappatura tra modelli concettuali (ontologie) e basi di dati (Ontology-based Data Access).

3 La piattaforma Isabox

3.1 Disegno della soluzione

L'architettura della piattaforma Isabox è stata ideata con l'obiettivo di fornire flessibilità evolutiva di configurazione, robustezza e sicurezza. Essa è concepita come un'infrastruttura da rendere disponibile nell'ambiente operativo del cliente (*on premise*) per essere naturalmente conforme ai requisiti di riservatezza e sicurezza. La piattaforma è struttrata per fornire servizi attraverso interfacce applicative a moduli di *front-end* che integrano l'aspetto di interazione con l'utente all'interno del contesto (portale, interfaccia grafica, applicazione) del cliente.

[1]AgID, Libro bianco sull'IA, https://ia.italia.it/assets/librobianco.pdf

[2]http://qallme.fbk.eu/

La piattaforma si articola su due livelli: (1) la gestione della sessione con l'utente, responsabile del flusso delle interazioni e proprietario delle informazioni sia di contesto sia di sessione (*episodic memory*) e (2) un sistema di servizi di base (*micro-services*) integrati, disponibili tramite API aperte (OpenAPI[3]). I micro-servizi rilsciati nella prima versione della piattaforma includono:

- Natural Language Processing (NLP)
  - Tokenization - POS Tagging - Lemmatization
  - Dependency parsing - Semantic Role Labelling
  - Short Sentence Similarity
- Knowledge Base
  - Conceptual Annotation
  - Frame Extraction
  - Conjuctive Query Answering
- Unstructured Question Answering
  - Retrieve and Rank
  - Question-Answer Pairs (FAQ)

.

3.2 Servizi di Natural Language Processing

I servizi di NLP si basano sulla piattaforma Open Source KotlinNLP[4]. Questa piattaforma implementa tecniche di Deep Learning allo stato dell'arte per riconoscere la struttura profonda di un testo [Grella e Cangialosi, 2018], ed è integrata con risorse linguistiche specifiche per l'italiano[5].

I moduli di analisi di KotlinNLP coprono l'intero "Stack NLP", dalla suddivisione di un testo in frasi e parole (tokenizzazione) alla lemmatizzazione e analisi morfologica (pos tagging), dall'analisi sintattica a dipendenze (dependency parsing) al riconoscimento dei ruoli semantici e delle tracce (i.e soggetti sottintesi o riferimenti pronominali) fino all'estrazione di entità nominali e temporali e categorizzazione automatica.

3.3 Servizi di Knowledge Base

La piattaforma Isagog adotta un approccio Ontology-based (OBDA)[Xiao *et al.*, 2018] per l'accesso alle basi di dati amministrative[6]. Questo consente di tenere separata la concettualizzazione (ontologia) dalle sorgenti di dati, a beneficio della generalità e della riusabilità dei modelli concettuali. Le interrogazioni (query) sono sempre formulate nell'alfabeto dell'ontologia, qualsiasi sia il database sottostante, essendo i livelli concettuale e logico messi in relazione mediante opportune regole di mappatura.

Per quanto concerne la redazione delle ontologie, si è adottato un approccio "fondazionale" basato su una *upper ontology* ispirata alle proposte di DOLCE [Gangemi *et al.*, 2002] e UFO [Guizzardi e Wagner, 2010]. I moduli di dominio includono al momento: Anagrafica, Documentazione, Organizzazione e Beni, per un totale di circa 200 tra classi e proprietà e 1200 assiomi di significato.

Queste ontologie, redatte in OWL2, saranno in futuro allineate con quelle sviluppate nell'ambito del Data Analytics Framework (DAF) di AgID[7]. Ad esse sarà affiancata una "ontologia di lessicalizzazione" contenente annotazioni che legano concetti, ruoli (proprietà) e istanze a elementi linguistici (lessemi ed eventuali vincoli morfo-sintattici) secondo uno specifico modello di interrelazione tra ontologia e lessico [Oltramari e Vetere, 2008].

3.4 Servizi di Question Answering

Molte informazioni di rilievo, specialmente per il personale delle Amministrazioni, è contenuto in documenti testuali, come delibere o determine. Inoltre, molte Amministrazioni sono in grado di predisporre coppie domanda-risposta (anche note come *Frequently Asked Questions* (FAQ)) in grado di soddisfare un gran numero di richieste dei cittadini, solitamente rivolte agli Uffici per le Relazioni col Pubblico (URP)[8].

Il servizio di QA non strutturato si baserà sulle funzionalità di Elasticsearch[9], opportunamente integrate con algoritmi di *similarity* e *ranking*. Il sistema risolverà le interrogazioni sia rispetto alle basi documentali, offrendo risposte contestuali (passaggi rilevanti), sia sulle coppie domanda-risposta eventualmente disponibili, ordinando le risposte in base alla confidenza.

3.5 Dialog Manager

Come sistema di dialogo, Isabox fa leva sulla pragmatica dell'interazione tipica del *question answering*, che rappresenta un caso particolare rispetto ai sistemi di assistenza virtuale o ai generici *chatbot*. Il dialogo consiste infatti nel chiarimento degli intenti dell'utente (*grounding*), nel caso in cui essi non siano completamente esplicitati nella *query* iniziale (*frame detection*).

Il sistema è in grado di trattare modelli di interrogazione (*frames*) di cui è possibile specificare le *slot*-chiave. Qualora uno di questi *frames* fosse istanziato (anche parzialmente) dalla domanda, il sistema guiderebbe alla suo completamento prima di inviarlo ai servizi di QA strutturato e testuale. Altrimenti, la domanda viene comunque concettualizzata (cioè interpretata nei termini delle ontologie di dominio) e una riformulazione in un linguaggio di query congiuntive (SPARQL) viene tentata. In tutti i casi, il servizio di query answering non strutturato (3.4) viene interrogato e le sue risposte (se reperite) valutate in termini di affidabilità.

Il comportamento *intelligente* del sistema nel suo complesso dipenderà dunque in larga misura dal modo in cui il gestore del dialogo interpreta i risultati dei servizi applicativi (rif. 3.1)

[3] openapi

[4] https://github.com/KotlinNLP

[5] Tra cui dizionari di "function words" https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2893 e "content words" https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2894

[6] L'implementazione prototipale attuale è basata su Ontop (https://ontop.inf.unibz.it/)

[7] https://docs.italia.it/italia/daf/

[8] Si veda il portale di Linea Amica: http://www.lineaamica.gov.it/

[9] https://www.elastic.co/

4 Conclusioni

Il rilascio della piattaforma Isabox è previsto nel Giugno del 2019. Successivamente al rilascio, saranno avviati alcuni progetti pilota che consentiranno di completare la copertura semantica del sistema e di valutare le prestazioni informative. Le ontologie generali e di dominio, assieme ai mapping e alla lessicalizzazione, verranno rilasciate con licenza aperta. Dal punto di vista della ricerca, la focalizzazione sarà sull'integrazione tra metodi di apprendimento automatico (modelli neurali) e metodi di rappresentazione della conoscenza ontologica e lessicale.

Riferimenti bibliografici

[Fernández, 2014] Raquel Fernández. Dialogue. In Oxford University Press, editor, *The Oxford Handbook of Computational Linguistics*, 2014.

[Gangemi *et al.*, 2002] Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, e Luc Schneider. Sweetening ontologies with dolce. In *EKAW*, 2002.

[Grella e Cangialosi, 2018] Matteo Grella e Simone Cangialosi. Non-projective dependency parsing via latent heads representation (LHR). *CoRR*, abs/1802.02116, 2018.

[Guizzardi e Wagner, 2010] Giancarlo Guizzardi e Gerd Wagner. Using the unified foundational ontology (ufo) as a foundation for general conceptual modeling languages. In *Theory and Applications of Ontology: Computer Applications*, pages 175–196, 03 2010.

[Magnini *et al.*, 2009] B. Magnini, M. Speranza, e V. Kumar. Towards interactive question answering: An ontology-based approach. In Proc. of the Workshop on Semantic Computing e Multimedia Systems (SCMS 2009), editors, *Third IEEE International Conference on Semantic Computing (ICSC 2009)*, 2009.

[Oltramari e Vetere, 2008] A. Oltramari e G. Vetere. Lexicon and ontology interplay in senso comune. In LREC 2008, editor, *Proc. OntoLex Workshop at 6th Intl. Conf. on Language Resources and Evaluation*, 2008.

[Soares e Parreiras, 2018] Marco Antonio Calijorne Soares e Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 2018.

[Webb e Webber, 2009] N. Webb e B. Webber. Special issue on interactive question answering: Introduction. *Natural Language Engineering*, 15(1):1–8, 2009.

[Xiao *et al.*, 2018] Guohui Xiao, Diego Calvanese, Roman Kontchakov, Domenico Lembo, Antonella Poggi, Riccardo Rosati, e Michael Zakharyaschev. Ontology-based data access: A survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5511–5519. International Joint Conferences on Artificial Intelligence Organization, 7 2018.